1-1-2013

Connectable Components for Protein Design

Gabriel B. Gonzalez
University of Pennsylvania, Gabriel439@gmail.com

Connectable Components for Protein Design

Abstract
Protein design requires reusable, trustworthy, and connectable parts in order to scale to complex challenges. The recent explosion of protein structures stored within the Protein Data Bank provides a wealth of small motifs we can harvest, but we still lack tools to combine them into larger proteins. Here I explore two approaches for connecting reusable protein components on two different length scales. On the atomic scale, I build an interactive search engine for connecting chemical fragments together. Protein fragments built using this search engine recapitulate native-like protein assemblies that can be integrated into existing protein scaffolds using backbone search engines such as MaDCaT. On the protein domain scale, I quantitatively dissect structural variations in two-component systems in order to extract general principles for engineering interfacial flexibility between modular four-helix bundles. These bundles exhibit large scissoring motions where helices move towards or away from the bundle axis and these motions propagate across domain boundaries. Together, these two approaches form the beginnings of a multiscale methodology for connecting reusable protein fragments where there is a constant interplay and feedback between design of atomic structure, secondary structure, and tertiary structure. Rapid iteration, visualization, and search glue these diverse length scales together into a cohesive whole.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Biochemistry & Molecular Biophysics

First Advisor
William F. DeGrado

Keywords
protein design, search engine, two-component systems

Subject Categories
Bioinformatics

CONNECTABLE COMPONENTS FOR PROTEIN DESIGN

Gabriel B. Gonzalez

A DISSERTATION

in

Biochemistry and Molecular Biophysics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2013

Supervisor of Dissertation

_____

William F. DeGrado, Professor, Pharmaceutical Chemistry

Graduate Group Chairperson

_____

Kathryn M. Ferguson, Associate Professor, Physiology

Dissertation Committee

Roland L. Dunbrack, Adjunct Professor, Biochemistry and Biophysics

Jeffery G. Saven, Associate Professor, Chemistry

Mark Goulian, Professor, Biology

Ravi Radhakrishnan, Associate Professor, Bioengineering

Li-San Wang, Associate Professor, Pathology & Laboratory Medicine

Andrej Sali, Professor, Bioengineering and Therapeutic Sciences

CONNECTABLE COMPONENTS FOR PROTEIN DESIGN

I dedicate my thesis to my dad, who instilled me with a love of science and taught me that the question is more important than the answer

ACKNOWLEDGMENTS

I owe my growth as a scientist to my friends in the DeGrado lab who taught me by their wonderful example. I really want to thank Brett Hannigan: my collaboration with you has been the most enjoyable and productive experience of my graduate study. I also am greatly indebted to my advisor, Bill DeGrado, who patiently mentored me until I could find my own path.

The Haskell community also has my gratitude. You all educated me and molded me into a better person. Paolo Capriotti deserves special mention: you first introduced me to the wonderful world of computer science and gave a voice to my vague and inarticulate thoughts.

I would like to thank my parents, Gabriel Gonzalez Sr., William (Mike) Scott, and Gabriella Scott, as well as my in-laws, Ba Phan and Khang Nguyen. You all taught me that there is no such thing as a self-made man. Without all of your support I would never have been able to complete my graduate studies.

To my wife Thao Phan: your love always lifted me up through difficult times. You always supported me, listened to me, and gave me a shoulder to cry on. You even moved across the country with me (twice!) to help me complete my degree. I will always love you and be forever indebted to you.

Finally, my kids, Michael and Gabriella, are my greatest source of joy and perspective. No matter how bad things got I could always come home to two sets of smiles and open arms.

ABSTRACT

CONNECTABLE COMPONENTS FOR PROTEIN DESIGN

Gabriel B. Gonzalez

William F. DeGrado

Protein design requires reusable, trustworthy, and connectable parts in order to scale to complex challenges. The recent explosion of protein structures stored within the Protein Data Bank provides a wealth of small motifs we can harvest, but we still lack tools to combine them into larger proteins. Here I explore two approaches for connecting reusable protein components on two different length scales. On the atomic scale, I build an interactive search engine for connecting chemical fragments together. Protein fragments built using this search engine recapitulate native-like protein assemblies that can be integrated into existing protein scaffolds using backbone search engines such as MaDCaT. On the protein domain scale, I quantitatively dissect structural variations in two-component systems in order to extract general principles for engineering interfacial flexibility between modular four-helix bundles. These bundles exhibit large scissoring motions where helices move towards or away from the bundle axis and these motions propagate across domain boundaries. Together, these two approaches form the beginnings of a multiscale methodology for connecting reusable protein fragments where there is a constant interplay and feedback between design of atomic structure, secondary structure, and tertiary structure. Rapid iteration, visualization, and search glue these diverse length scales together into a cohesive whole.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1 - Introduction

1.1 – Overview of protein design

Proteins can be likened to nature's microscopic robots, powering the majority of chemical and mechanical processes at the molecular level [4]. Nature's ubiquitous use of proteins testifies to their utility, and the better we can harness their power the more precisely we can control and orchestrate a wide variety of biological or chemical processes in exquisite detail.

There already exist several commercial applications of proteins, such as (A) transplanting existing natural proteins to new host organisms, such as in GMO food, (B) using proteins in a non-natural environment, such as textile processing [95], detergents [73], and biocatalysis [8], or (C) incorporating them into medical therapeutics, such as antibodies [13]. Additionally, several new commercial applications may emerge in the near future, including medical diagnostics [10], bio-ethanol production [64], vaccine delivery [84], drug delivery [102], and metabolic engineering [50].

The scientific state of the art has progressed even further. Many research groups have made great strides in designing large-scale super-molecular protein architectures, such as protein crystals [58] and symmetric polyhedra [53], switchable proteins [54], and highly potent catalysts [89]. However, as these efforts grow in complexity the reliability of the design process decreases, and the even successful and renowned research labs such as the David Baker group can go through tens of designs on their more challenging projects [28]. This presents an expensive and intimidating prospect for newcomers to the field who wish to break new scientific ground.

Unlike software programs, proteins are difficult to "debug" when things go wrong. A programmer can connect a software debugger and to a failed program to get a detailed portrait of the programs' internal state in order to diagnose problems. In contrast, a protein biochemist's diagnostic tools are more limited: they may run a gel and hypothesize why their protein expression product migrates at the wrong size, assuming that it expresses at all. As pitfalls accumulate it becomes increasingly difficult to systematically avoid them and we should devise new ways to stem the tide of "bugs" in order to improve the quality assurance of the protein design process.

1.2 – Rational and irrational protein design

Rational protein design is one approach that aims to solve the reliability problem that plagues protein design. This school of thought began as an attempt to understand the first principles underlying protein form and function [78] so that we can predict with confidence which designs will succeed and which designs will fail with less trial and error.

The opposite of rational design is "irrational design", which emphasizes large-scale exploration of an enormous number of potential solutions, the great majority of which are expected to fail. Directed evolution exemplifies this approach, where researchers generate large libraries of protein mutants and using a high-throughput screen or selection process to discriminate which mutants possess a functional property of interest [5]. However, I use a definition of irrational design which is intentionally broad to also include high-throughput computational screens [55] as well. Like directed evolution, these computational screens emphasize trial and error over understanding, although one can try to reverse engineer the numerous outcomes to uncover

new first principles for rational design. Both computational and in vitro screens can produce a large number of variations on a successful design which can provide detailed information about what role individual mutations play [101,106].

All irrational approaches share the same disadvantage: we cannot explore complex designs easily. For example, mutating ten positions within a protein chain to ten possible residues each requires screening 10^{10} possible combinations, which pushes the limits of phage display selections [91]. Even then, such large selections are not as desirable as lower throughput screens which can provide more accurate measurements of fitness, but at significantly reduced library sizes (typically at most 10^4 variants) [36]. Similarly, testing all of these variants computationally within a week would require that our selection algorithm must not take longer than 60 microseconds to run for each design we wish to test.

These limitations restrict brute-force searches to testing incremental changes to a protein rather than designing large pieces at a time. Designing on the ten-residue scale is appropriate for fine-grained details such as a protein's active site or a protein binding interface, but you cannot design a medium-sized protein of over 100 residues from scratch this way and you need an alternative approach to fill in the remaining bulk of the protein, either by reusing natural scaffolds [28] or by building new scaffolds *de novo* [57].

1.3 – Designability

My thesis explores an alternative approach to designing proteins that builds proteins by connecting "designable" protein components together. A designable protein fragment is defined as a structural element that is more tolerant of mutation or diverse structural contexts

and the concept of "designability" dates back to simple lattice models of proteins which showed that multiple diverse sequences would independently converge on the same structure [62]. We call such a recurring structural motif designable since we can select from many possible variations on the motif, making it more amenable to design. A testament to the designability of natural protein scaffolds is Dahiyat and Mayo's redesign of a zinc finger domain scaffold to a completely new sequence which still produced the same tertiary structure [20].

The growing size of the Protein Data Bank (PDB) and increases in computational power provide an opportunity to harvest designable building blocks from a large repository of deposited protein structures, currently numbering over 95,000 entries. Grigoryan et al. took this approach of reusing natural protein components when they designed a viral-like protein coating for nanotubes by incorporating designable helix-helix interactions using the MaDCaT search engine [39]. Previously, the reusable unit of protein design was an entire protein domain, but MaDCaT opened the door to reusing smaller interactions between secondary structure elements.

We can consider even smaller reusable units of design by studying conserved atomic-level motifs. The Erebus protein search engine [87] allows one to search for conserved atomic substructures in order to assess how abundant they are within nature, although this has not yet been applied towards protein design.

1.4 – Connecting motifs

Identifying designable structural elements does not suffice to solve the protein design problem. Each designable element is only a piece of the puzzle and we must provide a structured way to stitch them together into a complete protein.

Grigoryan et al. succeeded in designing their viral coating for carbon nanotubes by also layering three-fold symmetry on top of two designable helix-helix interactions to generate a complete six-helix bundle. However, this approach does not generalize to proteins that are non-symmetric or whose asymmetric unit is larger than a few designable motifs.

Similarly, Erebus cannot be easily used for designability purposes because there is no way to easily connect conserved small atomic substructures into a unified whole. This is even more problematic on the atomic scale because one cannot apply symmetry to an atomic-level motif to build an entire protein. Moreover, these designable atomic substructures have tighter geometric, chemical, electrostatic requirements than designable secondary structure interactions, which makes them more difficult to connect. A hydrogen bond distance in an atomic-level motif may vary by tenths of an Å [93], whereas a helix-helix interaction may vary in crossing distance by over 1 Å [97].

Additionally, we must also be able to combine multiple heterogeneous designable structural elements in order to generate novelty. If we restrict ourselves to only incorporating one or two designable interactions then we limit ourselves to plagiarizing existing proteins. Knowledge-based design cannot be really considered *de novo* design until it can weave together many disparate elements from unrelated protein structures.

I term this the "connectable protein design" problem: how to combine designable protein components on multiple length scales into a unified protein without steric clashes, chemical mismatches, or other geometric conflicts. Solving this problem would greatly generalize the applicability of knowledge-based design.

1.5 – Project Summary

My thesis approaches the connectable protein design problem by exploring two separate approaches to combining reusable protein components together with as few conflicts as possible. The first approach operates at the atomic scale and the second approach operates at the protein domain scale.

In Chapter 2, I solve connectability at the atomic level by creating an interactive workflow for piecing together designable atomic substructures from proteins. This workflow centers on an all-atom search engine that I built and integrated with molecular graphics software that allows users to interactively discover and incorporate these designable motifs into their protein blueprints.

In Chapter 3, I study two-component signal transduction systems which frequently mix and match a limited set of domains in diverse ways to generate novel signaling proteins. I use quantitative structural analysis to study the interfaces between these components and tease out the basis for their interfacial flexibility which permits such diverse inter-domain connections.

The primary novel contributions of this thesis are:

- An all-atom search engine that outperforms other search engines by over two orders of magnitude, built with technical innovations reusable by other search engines

- The first integration of a protein search engine with molecular graphics software, both for building search queries and visualizing search results

- The first interactive and connection-based protein design methodology that bridges atomic interactions to tertiary structure

- The identification of the structural basis for modularity and loose coupling for domains from two-component signal-transduction systems

CHAPTER 2 – A Real-Time All-Atom Structural Search Engine for Proteins

2.1 – Abstract

Protein designers use a wide variety of software tools for *de novo* design, yet their repertoire still lacks a fast and interactive all-atom search engine. To solve this, we have built the Suns program: a real-time, atomic search engine integrated into the PyMOL molecular visualization system. Users build atomic-level structural search queries within PyMOL and receive a stream of search results aligned to their query within milliseconds. This instant feedback cycle enables a new "designability"-inspired approach to protein design where the designer searches for and interactively incorporates native-like fragments from proven protein structures. We demonstrate the use of Suns to discover protein motifs, interactively build larger protein fragments and identify scaffolds compatible with hot-spot residues.

2.2 – Introduction

Protein structural bioinformatics rapidly approaches a big data crisis as the last decade has witnessed a dramatic increase in protein structure depositions. In 1993 researchers had just over 23,000 searchable structures at their disposal in the Protein Data Bank (PDB), while today we have over 95,000. This rapid structural expansion could inform protein design, structure determination, and structure prediction by providing numerous examples of native-like structural interactions in exquisite detail, but researchers lack high-powered computational tools to intelligently explore large structural data sets in detail.

One of the first popular protein structural search tools developed for this purpose was Dali by Holm and Sander [41]. Dali uses distance maps formed by calculating pairwise α-carbon

distances to form a two-dimensional representation of a three-dimensional protein. Regions of similarity between two distance maps correspond to similar substructures in their respective proteins. Holm and Sander used Dali to create the Families of Structurally Similar Proteins (FSSP) database [42], which aligns substructures across entries in the Protein Data Bank (PDB) to form families and subfamilies of common folds. Researchers commonly use Dali to compare protein folds and infer homology [23,77,81].

The more recent MaDCaT search program [105] also uses α-carbon distance maps to search for similar protein backbone arrangements. However, where Dali uses a heuristic approach to detect structural similarity, MaDCaT takes a query backbone structure or motif and finds globally optimal structural matches within an entire structural database. This approach makes MaDCaT ideal for finding the best matches to frequently occurring motifs. These "designable" motifs promise to be excellent design scaffolds, and MaDCaT applied this approach to design a viral-like protein coat for carbon nanotubes from designable interactions [39].

Both Dali and MaDCaT return results after a several minutes of searching. For greater speed, Shyu et. al. developed ProteinDBS [88] in order to provide the first real-time protein backbone search. They use image processing techniques to extract a set of features from α-carbon distance maps and organize their structural database into a tree, allowing quick traversal and parallelism during searches. These optimizations allow them to return search results nearly instantly, but they limit themselves to searching for backbone α-carbons.

We required an all-atom search engine to guide the protein design process, so that we could search for proteins with similar active sites or binding motifs, explore protein scaffolds that can host a specific motif, and discover atomic-scale supporting interactions.

The state of the art for all-atom search is Erebus [87], which permits all-atom rigid substructure searches, but this is insufficient for our design purposes because we desired an interactive search process. Several bottle-necks in the Erebus search workflow impede a fluid design process, including time-consuming assembly of search queries, long search delays, and a web interface for retrieving results.

A truly interactive search tool must remove every single one of these bottlenecks to bring the feedback loop down from minutes to seconds and permit users to rapidly explore multiple design alternatives iteratively in atomic detail. Improved speed and faster feedback lets researchers to ask more sophisticated questions, explore structures more intelligently, and use limited collaboration time more efficiently.

The Suns protein search engine makes it easy to search and browse a database of protein structures at the atomic level. To our knowledge, Suns is the first real-time all-atom structural search engine and also the first to integrate seamlessly into the popular molecular visualization program PyMOL, so that researchers to easily click on motifs of interest, click search, and view aligned results within a fraction of a second. We expect Suns to inform and guide protein design, modeling, and structure determination by lowering the entry barrier to structural search so that it becomes a staple of every structural biologist's toolbox rather than a tool limited to programmers.

2.3 – Design and Implementation

2.3.1 – Overview

Our structural search engine greatly resembles a web search engine, even though these two types of engines index different types of data: web search engines commonly index linear text strings whereas our search engine indexes three-dimensional protein structures. Despite these differences, we still borrow many principles from web search engines [11] to improve search speed:

1. Divide structures into structural "pages" (3-D volumes) analogous to web pages
2. Divide these "pages" into structural "words" (chemical motifs) analogous to textual words
3. Create a forward index that matches sets of structural words to structural pages
4. Perform slower and more accurate filters after the fast forward index lookup
5. Return only as many results as required to avoid unnecessary computation

2.3.2 – Forward Index

Web search engines derive much of their speed by preprocessing the data set using a forward index that matches words to web pages [11]. The search engine can then tokenize each query into words and consult the forward index to rapidly return all pages that contain every word in the user's search query. Protein search engines can copy this trick, but they must first decide what volume size corresponds to a "page" and what chemical motifs correspond to "words".

Two opposing considerations constrain the choice of page and word size. The forward index resolves pages solely by their word counts, so larger words and smaller pages lead to more unique word counts per page and improves the selectivity of the forward index. However, users prefer the exact opposite: smaller words and larger page sizes increase the power and flexibility of user search queries. Therefore, optimizing a structural search engine requires balancing user needs against the efficiency of the forward index.

We select a compromise suitable for atomic-level search queries: we restrict structural pages to cubes approximately 15 Å wide and we define structural words to be connected chemical substructures ranging from 2 atoms (a hydroxyl) to 9 atoms (an indole ring) (**Figure 1**). Our choice of page size assumes that larger structural patterns of interest can be reduced to a network of bridging local interactions below the 15 Å length scale. Similarly, our choice of word size assumes that users will accept a modest restriction on search queries to groups of chemical motifs instead of groups of atoms. Like web search engines, we permit searches for multiple disconnected words, allowing users to assemble complex queries from these simple chemical building blocks.

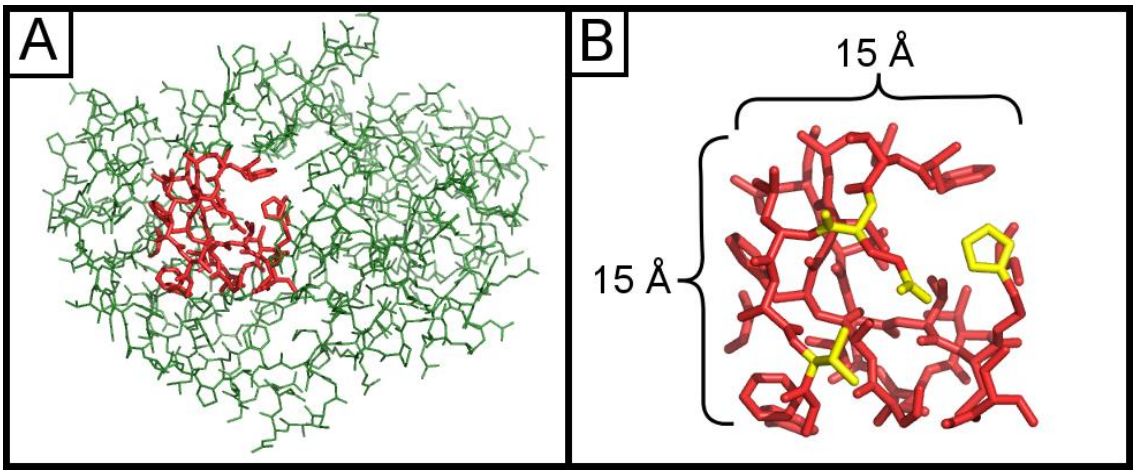

Figure 1 - Subdivision of protein structures. (A) An interior page highlighted in red from a protein of unknown function (PDB ID = 2FSQ), illustrating the maximum scale of search queries. (B) Example words (chemical motifs) within the same page highlighted in yellow. Pages are 15 Å x 15 Å x 15 Å cubes.

The forward index is a nested data structure and the outer level data structure is an array whose elements are word indices, one for each word that suns understands:

```
type PrimaryIndex = Vector WordIndex
```

Each word index contains a list of all matches, ordered by number of matches:

```
type WordIndex = Vector Matches
type Matches = Set PageID
```

The first element in the `WordIndex` vector consists of all pages that contain exactly one occurrence of the given word. The second element consists of all pages that match the motif twice, and so forth.

Pages are grouped by number of matches so that Suns can rapidly eliminate pages that do not have a sufficient number of matches. If the user searches for three carboxylic acids, then the search engine can immediately skip the first two elements of the `WordIndex` vector for carboxylic acids since they will be sets with fewer than three matches to carboxylic acids. Then it folds all the remaining elements using set union to retrieve all pages with at least three matches.

However, users can search for mixtures of diverse words, so to do this efficiently we query each word type independently and then take the intersection of all queries. So if the user requests two hydroxyls and two carboxylates, then the search engine will split this into two subqueries. First, it will compute the set of all pages with at least two hydroxyls and then compute the set of all pages with at least two carboxylates. Computing the intersection of these two sets identifies pages with that simultaneously contain at least two hydroxyls and at least two carboxylates.

2.3.3 – Structural pages

Suns partitions protein structures spatially into pages which non-overlapping cubes approximately 15 Å wide. Search results must fit within one of these pages, meaning that the search engine does not return search results that span more than one page. Non-overlapping pages were chosen for efficiency reasons, since overlapping pages would require an additional search step to remove duplicate search results contained entirely within overlapping regions.

Protein atoms are partitioned into buckets by using a truncated Morton code [71]. First, the X, Y, and Z coordinates are converted from floating point numbers to 21-bit integers using the following formula:

$$toInt(v) = \left\lfloor \frac{2^{21}(v - v_{min})}{\sqrt{2}(v_{max} - v_{min})} \right\rfloor$$

$$v_{max} = 9999.999$$

$$v_{min} = -999.999$$

v_{max} and v_{min} are the upper and lower bounds, respectively, for X, Y, and Z coordinates in the Protein Data Bank file format. $toInt$ rescales every floating point coordinate in this range to a 21 bit integer. The $\sqrt{2}$ in the denominator is a fudge factor to adjust the final page sizes to be approximately 15 Å (the true dimension is 15.19 Å).

An atom's X, Y, and Z coordinates are combined into a 63 bit integer using bitwise interleaving of their binary integer representations, which corresponds to a Morton encoding of the three coordinates. The index then assigns each atom to page by taking its Morton-encoded coordinate (the 63 bit integer) and dropping the 33 least significant bits. The remaining 30 bits correspond to the atom's page ID and multiple atoms can map to the same page ID because of truncation. Truncating the Morton encoded coordinates has the effect of portioning atoms into cubes 15 Å wide that tile space.

2.3.4 – Structural Words

We specify structural words using PDB files, which contain the specific residue and atom types to match. For example, one structural word consists of a single PDB file containing the Cα-Cβ-Cγ linker of phenylalanine. When users search for the three-carbons in phenylalanine's linker, their searches will not match tyrosine's linker, nor will they match three connected ring carbons

within a phenylalanine. This allows the search index to optionally resolve motifs that are otherwise chemically identical [15].

Structural words may also match more than one protein element, and in those cases we use multiple PDB files to specify the structural word: one PDB file per matching chemical motif. For example, one motif we index is a carboxylate, specified using two PDB files: one for glutamate's carboxylate and another for aspartate's carboxylate. User search queries for carboxylates will match either of these two groups.

The choice of structural words is customizable and for our public-facing server we select a default set of substructures appropriate for general-purpose searches (**Table 1**). The most important searchable substructure matches the four backbone atoms for any protein residue, which permits geometrically exquisite backbone searches that specify all backbone atoms and torsion angles. We partition flexible residues such as lysine and methionine into two separate words, and also isolate important chemical moieties into their own words, such as imidazole and guanidinium groups. Some chemical moieties are shared between residues, such as the hydroxyl group, which matches serine, threonine, and tyrosine. However, every residue except glycine possesses at least one unique structural word so that users can restrict searches to a specific residue.

2.3.5 – Tokenizing Words

The search engine must tokenize protein structures into words in two separate locations. First, the search engine must tokenizes the entire structural database into words since all search queries are specified in terms of words, not atoms. Second, the search engine must tokenize

every incoming search request into words before it can retrieve matching words from the database.

Before tokenizing, Suns converts protein structures to undirected graphs, with one graph per page in the structure. Graph nodes are atoms and vertices are bonds, so graphs are sparsely connected since the maximum degree of any node is 4 (the maximum number of bonds per atom). Also, since graphs are limited to atoms that fit within a single page they are small (fewer than 100 nodes/atoms).

The tokenization algorithm is implemented as a monadic, recursive descent backtracking parser [18], so that it can use Haskell's `do` notation as syntactic sugar for assembling parsing computations monadically. Parsers are conventionally associated with tokenizing text, but monadic parsers in Haskell generalize to other data structures, such as chemical graphs, and Suns takes advantage of this generalization to simplify graph tokenization code.

The monad used for parsing is a backtracking list monad enriched with state local to each branch of the search tree [47]:

```
newtype ParseS a =
    ParseS { unParseS :: StateT Structure [] a }
    deriving (Monad, Alternative)
```

The primitive parser takes two atom names as arguments, each of which uniquely identifies an atom within a specific residue type. This parser matches any bond that bridges two such atoms,

removing the matched bond from the graph and returning the indices of the nodes that matched:

```
pBond :: AtomName -> AtomName -> ParseS (Int, Int)
```

If more than one bond matches then `pBond` branches one search path per potential solution. If no bond matches then the current search path terminates and backtracks to try another potential solution. Note that removing the matched bond does not interfere with other branches of the backtracking search because each branch of the search maintains a separate copy of local search state.

The higher-level `pMotif` function builds on top of `pBond`, taking a motif graph as input and returning indices of a matching motif while simultaneously removing the matched motif from the protein's graph. `pMotif` invokes `pBond` as a subroutine once per bond found within the motif graph, incrementally checking that the connectivity of each newly matched bond is consistent with previously matched bonds:

```
pMotif :: Structure -> ParseS (Vector Int)
```

Like `pBond`, `pMotif` branches for every possible solution and backtracks if no solutions are found.

The `evalParseS` function runs any parser (including bond parsers, motif parsers, or further derived parsers), converting the parser to a list of potential solutions:

```
evalParseS :: ParseS a -> Structure -> [a]
```

This list of potential solutions is generated lazily [44], meaning that the algorithm only searches for as many solutions as we request, performing the minimal amount of computation necessary. Since Suns only uses the first solution the parsing algorithm usually does not explore all possible branches and defers unnecessary evaluation.

2.3.6 – Database

Our forward index is formally a *record level* inverted index, meaning that it only returns matches to pages, not to specific structural words within those pages. We supplement the forward index with a custom in-memory database that stores two pieces of information necessary to complete the search. First, the database stores correspondences between words in the forward index and atoms in each structural page. Second, the database also keeps compact representations of every structural page suitable for returning as search results.

The database is a single in memory nested data structure, where the top-level data structure is a vector with one element per page in the data set. Each element contains the PDB ID code for the structure the page originated from, a vector of atoms within that page, and then a nested data structure containing all words found in that page:

```
type SecondaryIndex = Vector (PDB, Vector Atom, Matches)
```

Matches are organized by words, words are organized by incidence, and each incidence is a list of integer indices into the page's vector of atoms indicating which atoms that word comprises:

```
type Matches   = Vector Words
type Words     = Vector Incidence
type Incidence = Vector Int
```

When the forward index produces a matched page, the secondary index remembers which atoms in that page correspond to the words advertised in the forward index. Sometimes the page contains more instances of a given word than the user required, such as when the user searches for two peptide bonds, and the page contains five. The page must try out every valid permutation of words that match the user's query, and the forward index minimizes the number of permutations by prioritizing pages that most closely match or exactly match the minimum required word count.

2.3.7 – Alignment and RMSD

Suns uses the Kabsch algorithm [48] to rapidly align each permutation to the user's search query. The Kabsch algorithm requires an exact atom-for-atom correspondence between the user's search query and a candidate motif, and Suns compiles this correspondence from precomputed atomic correspondences for each stored motif in the custom database. After alignment, the search engine only returns search results that match the search query within a specified root-mean-square deviation (RMSD) cutoff.

For each result below the RMSD cutoff, Suns aligns the matching page to the search query and return the page as the search result. If a page contains multiple matches Suns aligns each match separately and returns them as separate results. This superimposes every search result and context on the original query for ease of visual comparison and downstream post-processing.

2.3.8 – Streaming Results

The search engine does no global ranking of results by RMSD. This means that the search engine will immediately stream any result within the specified RMSD cutoff to the user, which allows the user to begin visualizing results before the search has completed, improving interactivity.

Additionally, the search protocol requires the user to specify the number of desired results up front. While the user may request an unlimited number of results in theory, in practice the search clients default to 100 search results, similar to how a web search engine will default to 10 search results. This allows the search engine to stop processing the request after supplying the specified number of results, which reduces server load. Also, the search engine may also optionally specify a search timeout to further reduce server load for users that request a large number of search results.

2.3.9 – Data Set

The public search engine uses PISCES [99] as the non-redundant protein structure data set, selecting a 20% sequence identity, 1.6 Å resolution, and 0.25 R-factor cutoff, which currently corresponds to 2058 chains. The search engine's available memory limits how many structures

it can index, and our stress tests on the largest PISCES data set (90% identity, 3.0 Å, 1.0 R-factor cutoff, 24,218 chains) required 89 GB of memory or 1 GB of memory per 272 protein chains.

2.4 – Results

2.4.1 – Building Motifs

Suns lets users explore the "designable" space of protein motifs by expanding on small initial fragments, such as building a helix N-terminal capping motif beginning from a single guanidinium group. One might begin by searching on the guanidinium fragment from an arginine, which recruits a cluster of nearby carboxylates forming a salt bridge with the arginine (**Figure 2A**). Adding one of these carboxylates to the search query refines the motif further, revealing a preferred rotamer for the arginine when interacting with a carboxylic acid (**Figure 2B**), and adding a preferred rotamer to the search query crystallizes a complete N-terminal capping motif (**Figure 2C**).

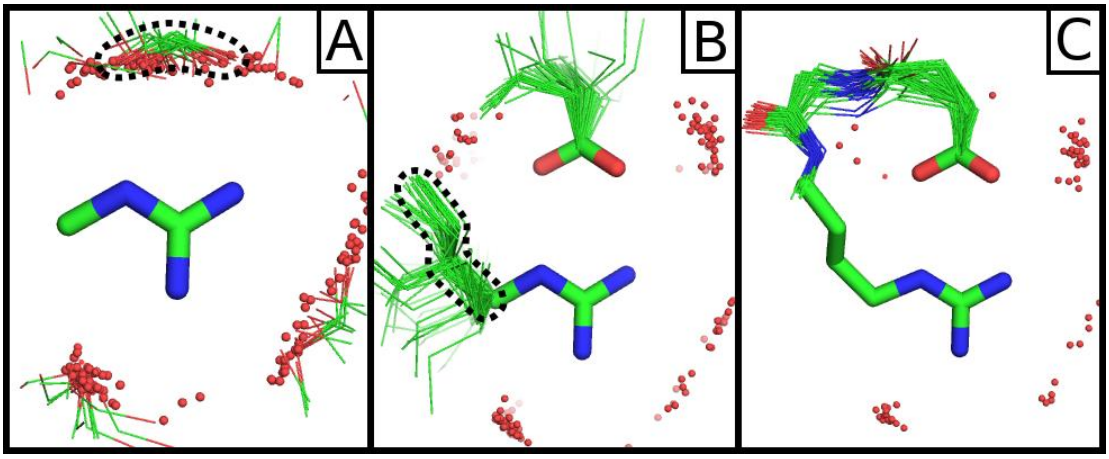

Figure 2 - Incremental assembly of a motif. (A) An initial search for a guanidinium fragment reveals a cluster of nearby carboxylates. (B) Refining the search with one carboxylate from the results reveals a specific linker preference for both the aspartate and arginine involved in the salt bridge. (C) Adding the most common linker for arginine and repeating the search reveals

that this salt bridge is part of an N-terminal capping motif. Search queries are represented as thick sticks and search results are shown as thin lines. Dashed lines highlight clusters in the search results, which are filtered to show the specific residue fragments of interest and neighboring water molecules within 3.0 Å as red spheres. Search parameters and fragments listed in **Table 2**.

The large number of close geometric matches to the final search query suggests that this is a highly "designable" motif. Incremental searching allows users to rapidly explore and prototype designable native-like interactions like these with very little prior knowledge in protein folding or biophysics. Moreover, a user can discover the motif by gradually refining a specification rather than specifying all the necessary interactions up front. This benefits people who may not even know what designable interactions look like and simply wish to explore what options they have available.

2.4.2 – Assembling Larger Fragments

Users can build tertiary interactions for proteins as well. To demonstrate this, we search for a valine from glucose binding protein and grow that into three small β strands with three residues per strand.

Beginning from an interior valine from glucose-binding protein, we seed the two adjacent β strands with highly populated residue clusters on each side corresponding to a valine and tyrosine (**Figure 3A**). To grow the three β strands in both directions, we search for pairs of residues at a time to identify new clusters of residues within the search results that we can insert into the sheet (**Figure 3B**). The PyMOL search client permits a qualitative inspection of residue preference at selected positions by cycling through visualizing each residue type. This

process not only provides a rough measure of residue preference, but also reveals rotameric preference, the kind of detailed information that a sequence logo would not reveal.

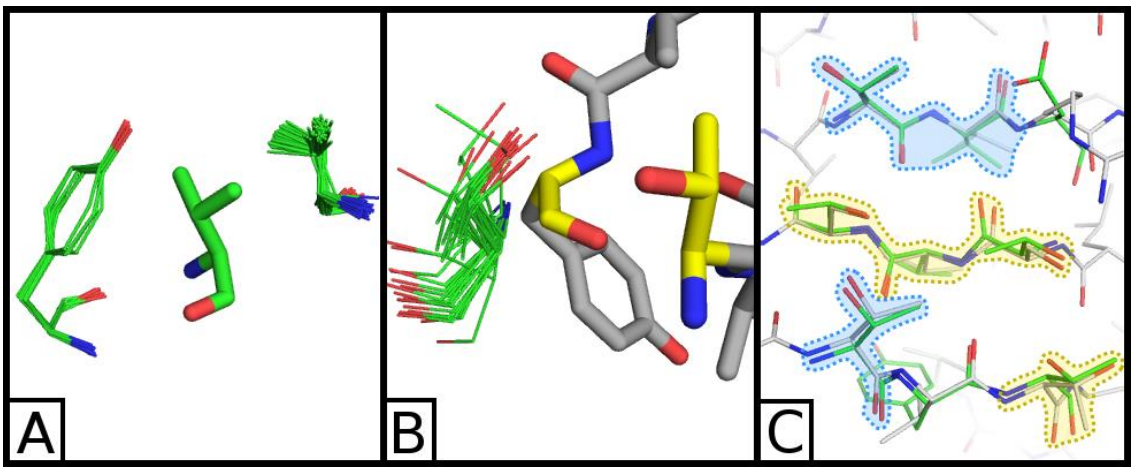

Figure 3 – Building a tertiary interaction.(A) Three strands are seeded by searching on a valine, which reveals two nearby clusters of valine and tyrosine. (B) Strands are extended one residue in each direction by searching for pairs of residues (colored yellow), yielding clusters of potential inserts (colored green). (C) The final backbone fragment (green) is fed to MadCaT, which identifies multiple compatible scaffolds. One such scaffold (PDB ID=1E54, colored light grey) possesses many exact residue/rotamer matches to the assembled fragment (blue highlights) and many close matches (yellow highlights) that differ by a related residue (threonine to serine or valine to isoleucine) or by varying the rotamer.

We repeat this process of iteratively searching for pairs of residues at a time and incorporating clusters from the search results until we assemble a native-like fragment of a sheet where almost every residue originates from a unique protein structure (two disconnected threonines were inadvertently drawn from the same structure). This then provides α-carbon coordinates that we feed into the backbone search engine MaDCaT [39], which finds suitable scaffolds to incorporate this fragment. One MaDCaT search result greatly resembles the β sheet built using Suns (**Figure 3C**). This illustrates how the local search capabilities of the Suns search engine

complement existing coarse-grained search tools by bridging the gap between the world of smaller atomic interactions and the world of larger secondary-structure interactions.

2.4.3 – Connecting Hot Spot Residues

Suns can also be used to find scaffolds compatible with specified residues to provide an alternative implementation of the hotspot residue approach to design [28]. The user can select the hotspot of interest within PyMOL, search, and find all proteins in the PDB that position the given hot spot residues in the specified geometry.

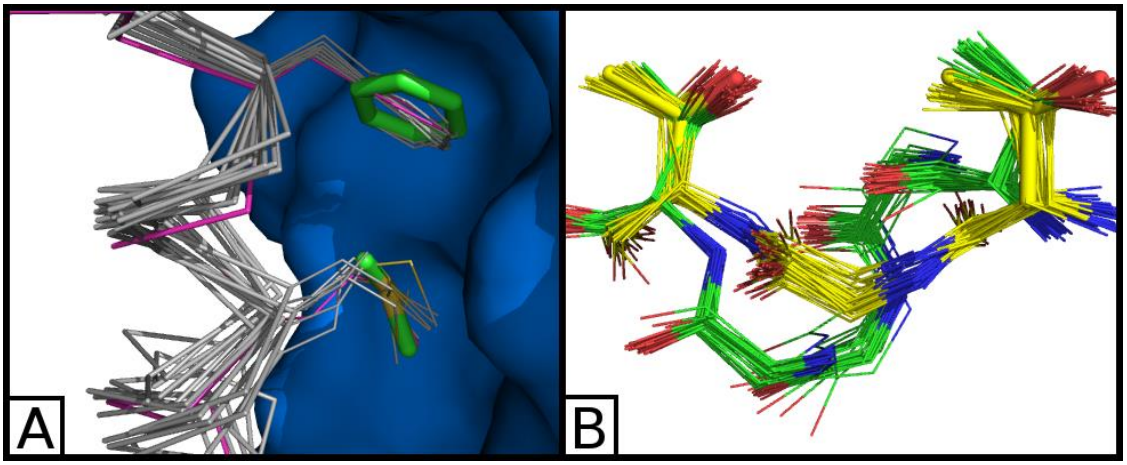

Figure 4 - Finding backbones compatible with hot spot residues. (A) A Suns search at 0.7 Å RMSD cutoff for two hotspot residues previously identified by RosettaDock [38] for a hemagglutinin binder [28]. The majority of search results are helices that closely match the final designed protein. The search query is shown in thick green sticks, the search result matches are shown as grey α-carbon traces, and the designed hemagglutinin binder is shown as a purple α-carbon trace against a blue hemagglutinin surface. (B) Searching for two threonine side chains at 0.6 Å RMSD cutoff reveals two backbone clusters that can connect them, one corresponding to an α helix (green) and the other corresponding to a β sheet (yellow). The original search query is shown in thick yellow sticks.

For example, Suns recapitulates the local backbone of a designed hemagglutinin binder [28].

Figure 4A illustrates how searching for fragments of the original hotspot residues reveals a

prominent cluster of α helices matching the designed protein structure, indicating that the secondary structure of the interface could have been predicted solely from designability.

Not every hotspot search will return a single solution for the backbone. Sometimes searching for disembodied residues will reveal multiple distinct ways to thread the backbone between them (**Figure 4B**).

2.5 – Discussion

2.5.1 – User friendliness

Suns greatly improves on existing search engines in terms of ease of use. This encourages use among a broader scientific audience, particularly people who are not programmers. Reducing the "activation barrier" to protein search encourages users to apply protein search in novel and previously unanticipated ways that may have never materialized had they been limited by the availability of collaborations to computational researchers.

Ease of use also benefits non-scientists or students, who can now enjoy a private and unfiltered learning and discovery process. Because of the low entry barrier, Suns can also be used as a teaching tool within the classroom to present general principles of protein biophysics. An instructor can show how electron donors cluster around the ε-nitrogen of a tryptophan, or how water molecules form hydrogen bonds with the helical backbone on the soluble face of a helix.

After all, it is one thing to be told that a structural element is a commonly recurring motif and it is another thing entirely to see with one's own eyes 100 real examples of that motif from the PDB all superimposed on top of one another.

2.5.2 – Speed

The Suns search engine greatly advances the state of the art in atomic substructure search speed. These optimizations are reusable by other search engines, such as grouping elements into logical units so that a forward index can be employed or partitioning searchable spaces into local volumes to prevent combinatorial explosion of atomic configurations.

This speed comes at a price: the most important optimization proved to be the coarse-graining of atomic substructures into groups of atoms corresponding to chemical motifs. Suns differs from the Erebus search engine by not permitting searches for arbitrary atomic configurations and instead only allows searches for collections of motifs. This motif-based approach allows Suns to improve the efficiency of its forward index, since motifs are more unique than atoms. For example, the indole ring of a tryptophan ring is highly unique thanks to the rarity of tryptophan, which allows the forward index to skip large volumes of protein structure that lack tryptophan. On the other hand, if you view the indole ring as a nondescript bag of atoms (8 carbons and a nitrogen) then this uniqueness is lost and the forward index cannot eliminate many structural regions.

A useful avenue for inquiry would be to combine the best features of both Suns and Erebus. It may be possible to let the user search for atomic subsets of chemical motifs, but under the hood the search engine supplies the entire motif to the forward index for the purpose of eliminating potential results. If this worked, then it would combine the atomic granularity of Erebus with the search speed of Suns.

2.5.3 – Potential Applications

Suns was originally built for protein design, but might prove useful to structural biologists. They may be able to use Suns as a generalized PROCHECK [59] that can quickly assess if a given structural element was modeled accurately or not.

2.5.4 – Generalizing protein search

We initially built Suns to guide the protein design process, but we are releasing it as a general purpose search engine so that others may reuse it for applications we did not previously anticipate.

Currently the public search engine only indexes protein structures. We also plan to add support for ligand search queries so that Suns can be used for drug design. While this paper describes a protein-specific application of the search engine, the underlying algorithm can be readily generalized to ligands and other macromolecules. Such a generalized search could prove useful for drug discovery.

2.6 – External resources

The official web site for Suns:

http://www.degradolab.org/suns/

PyMOL plugin – Master branch:

https://github.com/godotgildor/Suns

PyMOL plugin – Version referenced in manuscript:

https://github.com/godotgildor/Suns/commit/eed6b183097b6afb93c5336fb508f461eb9c9a8c

Command line search tool – Master branch:

https://github.com/Gabriel439/suns-cmd

Command line search tool – Version referenced in manuscript:

https://github.com/Gabriel439/suns-cmd/commit/92c37b07b86e7e3136f732709eade5acb960adf0

Search engine – Master branch:

https://github.com/Gabriel439/suns-search

Search engine – Version referenced in manuscript:

https://github.com/Gabriel439/suns-search/commit/1100a3c12a34d1ba92f2531a4c1fdea0bb2339f5

2.7 – Acknowledgments

Brett Hannigan and William F. DeGrado are co-authors on this manuscript. Brett Hannigan developed the PyMOL plugin for Suns. Gabriel Gonzalez built the search engine backend. William F. DeGrado contributed to the project's conception. All authors were involved in writing the manuscript.

2.8 – Supporting Information

Table 1 - Default Motif Set. Default motifs indexed by the public server hosted at suns.degradolab.org. (Motif Name): The common name for the motif. (Residue and Atom Names): The atom names used to define the motif. Some motifs may match multiple residue types, in which case all matching residues are listed with their corresponding atom names.

Motif Name	Residue and Atom Names
Alanine	Ala(Cα,Cβ)

Arginine Linker	Arg(Cα,Cβ,Cγ,Cδ)
Asparagine Linker	Asn(Cα,Cβ,Cγ)
Aspartate Linker	Asp(Cα,Cβ,Cγ)
Carboxamide	Asn(Cγ,Oδ,Nδ), Gln(Cδ,Oε,Nε)
Carboxyl	Asp(Cγ,Oδ1,Oδ2), Glu(Cδ,Oε1,Oε2)
Cysteine	Cys(Cα,Cβ,Sγ)
Glutamine Linker	Gln(Cα,Cβ,Cγ,Cδ)
Glutamate Linker	Glu(Cα,Cβ,Cγ,Cδ)
Guanidinium	Arg(Cδ,Nε,Cζ,Nη1,Nη2)
Histidine Linker	His(Cα,Cβ,Cγ)
Hydroxyl	Ser(Cβ,Oγ), Thr(Cβ,Oγ), Tyr(Cζ,Oη)
Imidazole	His(Cγ,Cδ,Nδ,Cε,Nε)
Indole	Trp(Cγ,Cδ1,Cδ2,Cε1,Cε2,Nε,Cζ1,Cζ2,Cη)

Isoleucine	Ile(Cα,Cβ,Cγ1,Cγ2,δ)
Lysine End	Lys(Cδ,Cε,Nζ)
Lysine Linker	Lys(Cα,Cβ,Cγ,Cδ)
Methionine End	Met(Cγ,Sδ,Cε)
Methionine Linker	Met(Cα,Cβ,Cγ)
Peptide Bond	All Residues(Cα,C,N,O)
Phenylalanine Linker	Phe(Cα,Cβ,Cγ)
Phenyl	Phe(Cγ,Cδ1,Cδ2,Cε1,Cε2,Cζ), Tyr(Cγ,Cδ1,Cδ2,Cε1,Cε2,Cζ)
Proline Ring	Pro(Cβ,Cγ,Cδ)
Serine Linker	Ser(Cα,Cβ)
Threonine Linker	Thr(Cα,Cβ,Cγ)
Tryptophan Linker	Trp(Cα,Cβ,Cγ)
Tyrosine Linker	Tyr(Cα,Cβ,Cγ)

Valine	Val(Cα,Cβ,Cγ1,Cγ2)

Table 2 - Search Parameters for all figures. (Figure): The figure and sub-figure the selections and searches correspond to. (Selection / {Search}): No braces indicates a saved selection referenced by searches. Braces indicate a search based in terms of previous selections of the form {sel1, sel2, …}. "sc" indicates only the side-chain was taken from the previously saved selection and "bb" indicates only the backbone atoms were used. (Structure): The PDB ID the selection originated from. (Result ID): The search result serial ID number to disambiguate selections where there are multiple results from the same PDB ID. (Chain): Chain the selection originated from. (Residue): Residue selected. (Atoms): Selected atoms. (RMSD Cutoff): Root-mean-squared deviation cutoff used for a given search. With the exception of initial selections for each figure, all selections are derived from results returned from the preceding search query in the table. †: Structure provided by the David Baker laboratory for their hot spot motif for the hemagglutinin binder [28].

Figure	Selection / {Search}	Structure	Result ID	Chain	Residue	Atoms	RMSD Cutoff (Å)
2	1	2GBP	N/A	A	Arg4	Cδ,Nε,Cζ,Nη1, Nη2	
	{1}						0.2
	2	3A6R	1	A	Asp61	Cγ,Oδ1,Oδ2	
	{1,2}						0.2
	3	3P02	0	A	Arg325	Cα,Cβ,Cγ,Cδ	
	{1,2,3}						0.3

3A	4	2GBP	N/A	A	Val88	Entire Residue	
	{4}						0.1
	5	4ASM	0	B	Val353	Entire Residue	
	6	2WUR	0	A	Tyr92	Entire Residue	
3B	{4bb,6bb}						0.2
	7	2JCQ	1	A	Thr151	Entire Residue	
	{4bb,7}						0.2
	8	2JCQ	0	A	Thr149	Entire Residue	
	{7sc,8bb}						0.5
	9	3B34	0	A	Thr37	Entire Residue	
	{5bb,8sc}						0.5
	10	3SUU	0	A	Asp102	Entire Residue	

	{6bb,7sc}						0.5
	11	3D9A	0	H	Thr482	Entire Residue	
	{6bb,8sc}						0.5
	12	3Q1I	0	A	Thr561	Entire Residue	
4A	13	†	N/A	B	Met503	Cγ,Sδ,Cε	
	14	†	N/A	B	Phe504	Cγ,Cδ1,Cδ2,Cε1,Cε2,Cζ	
	{13,14}						0.7
4B	{7sc,8sc}						0.6

CHAPTER 3 – PhoQ, a histidine kinase, signals across the membrane using a scissoring mechanism

3.1 – Abstract

All organisms signal across membranes to sense and adapt to external environments. Bacteria signal across the membrane primarily using two-component systems (TCSs), consisting of a membrane-spanning sensor histidine kinase and a cytoplasmic response regulator. In *Salmonella enterica* and other gram-negative bacteria, the PhoPQ TCS aids virulence by sensing cations, antimicrobial peptides, and low pH, yet little is known about what structural changes transmit the signal across the membrane. Here, we built a model of PhoQ signal transduction using Bayesian inference, based on disulfide crosslinking data and homologous crystal structures. We conclude that PhoQ inhabits two structurally distinct states that alternate via a scissoring motion. These states differ in regions critical to signal transduction such as the membrane depth of the sensor's acidic patch and the helical packing of the dimer interface. A comprehensive structural comparison of homologous two-component domains indicates this scissoring transition also occurs in other TCSs, suggesting a general mechanism of signal transduction.

3.2 – Introduction

The PhoQ sensor histidine kinase belongs to a family of two-component signal transduction systems, which dominate signaling across prokaryotic membranes [92]. These systems respond to diverse environmental signals, such as low pH [32], small molecules [49,60], ions [33], and peptides [52], and regulate critical responses, such as ion transport and virulence [69]. A prototypical two-component system (TCS) includes a transmembrane sensor histidine kinase

(HK) and a cytoplasmic response regulator [65]. The periplasmic sensor responds to environmental signals by promoting autophosphorylation of a conserved histidine, followed by phosphotransfer to a conserved aspartate residue on its corresponding cytoplasmic response regulator. Phosphotransfer activates the response regulator, which in turn modulates genetic response [80].

Although TCSs have been shown to be diverse [25], the topology of a canonical sensor HK (**Figure 5A**) consists of a periplasmic sensing domain flanked by two transmembrane (TM) helices, followed by one or more small domains, such as HAMP in PhoQ (named for the domain's presence in *hi*stidine kinases, *a*denylyl cyclases, *m*ethyl-accepting chemotaxis proteins, and *p*hosphatases) [30], and finally the kinase domain. This domain is typically known as the *d*imerization and *h*istidine *p*hosphotransfer domain (DHp), which contains the substrate (a conserved histidine) for autophosphorylation. The second part of this domain is a catalytic, ATP-binding domain that mediates autophosphorylation and phosphotransfer reactions. A functional histidine kinase is homodimeric (**Figure 5B**) with an extended dimer interface along the entire length of the molecule [35]. TCSs frequently reuse these domains, so mechanistic insights into PhoQ inform TCS signal transduction in general.

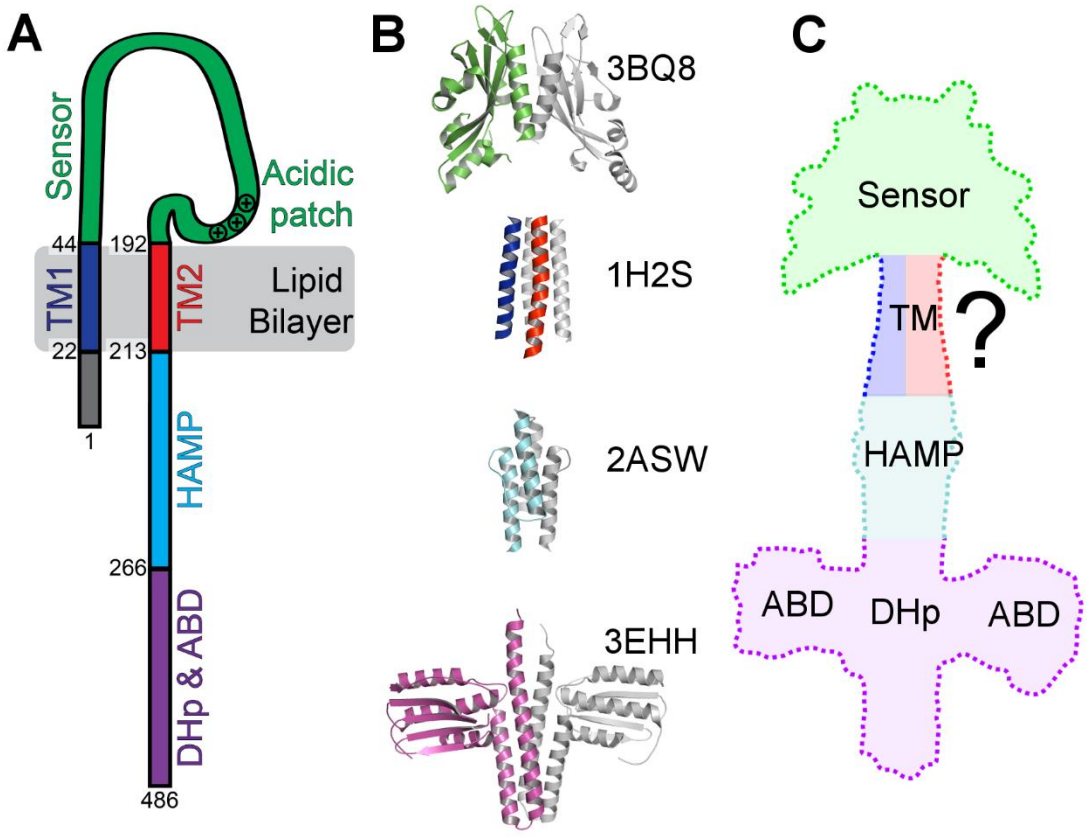

Figure 5 - Structural representations of PhoQ. (A) Schematic of the topology of a PhoQ monomer. The numbers indicate residue numbers for *E. coli* PhoQ (UniProt: P23837). (B) Crystal structures used for structural comparison of each domain of PhoQ. The corresponding PDB ID is listed next to the structure. One monomer is color-coded and the other monomer is in grey. (C) A model of the first three domains of PhoQ: sensor, transmembrane (TM), and HAMP domains. The dimerization and histidine phosphotransfer domain (DHp) and ATP-binding domain (ABD) are added for clarity but were not modeled.

Structural efforts have attempted to elucidate the mechanistic details of signal transduction spanning several domains from the periplasmic sensors to the cytoplasmic DHp domain, and several structures have been reported. Crystal structures are now available for multiple domains of two-component and chemotaxis systems [2,22,27,103], including a structure of the periplasmic sensor domain of PhoQ [17]. NMR and X-ray structures have also been solved for HAMP domains as well as transmembrane regions [24,82]. Recently, a full length structure of an

engineered, cytoplasmic two-component sensor (lacking a TM domain) was determined [22], and the structure of the cytoplasmic region of VicK, from *Streptococcus mutans*, was reported [98]. Despite these advances, there remain several competing proposals for a unified mechanism of transmembrane signaling.

Early studies on chemotaxis systems proposed a piston-like mechanism for signal transduction based on cysteine-scanning disulfide formation, mutagenesis, and crystallography [16]. In this model, a transmembrane helix signals across the membrane using a rigid translation orthogonal to the plane of the membrane [26], and later structural comparisons of the TorS TCS supported this hypothesis [70]. However, the measured displacements in the piston model are quite small in comparison to the length of the sensor HK protein itself or to the conformational changes expected to power a large rearrangement of the catalytic site.

In contrast, studies on cytoplasmic signaling domains propose a gearbox model where helical rotations within a four-helix bundle change the packing interfaces between helices [45]. Other observed motions include inter-helical torqueing [22], helix-bending [98], or DHp domain cracking [19]. Another study posits a combination of these models [14]. However, all proposed mechanisms lack a crucial ingredient: structural evidence for these motions extending into the transmembrane domain.

Critical to a membrane signal transduction model is a structural model of the TM portions of sensor HKs. Three structures of monomeric HK transmembrane domains were recently solved using NMR of isolated domains in micelles [66]. All three of the reported structures are limited in their utility for modeling a physiological dimeric interface, and without structural analyses

from a *bona fide* HK TM domain the structural starting point is not obvious. However, one crystal structure has been solved for the dimeric TM domain of a homologous protein: the HtrII sensory transducer [37]. A previous study utilized the HtrII X-ray structure as a model for the transmembrane domain in HKs [37], and we have also reported similarities between the TM domains of HtrII and PhoQ [34]. We demonstrated that the same pronounced water hemi-channel observed in HtrII plays an important mechanistic role within PhoQ [34].

Previously, we explored local changes in the TM domain by combining molecular dynamics simulations with disulfide crosslinking data [61]. To elucidate larger scale changes across the membrane, we incorporate new crosslinking data in the HAMP and juxtamembrane regions of PhoQ with previous data, and analyze it using multi-state Bayesian modeling [9,79]. This approach provides the first investigation into the structures of the two signaling states of PhoQ, which interconvert through a large scissoring motion. Our subsequent quantitative structural analysis of additional TCSs also divulge large and recurring scissoring motions. Scissoring accounts for a greater proportion of observed motions than proposed piston-shift [16] or gearbox [24] signal transduction mechanisms.

3.3 – Results

We probed the TM domain and the neighboring HAMP and periplasmic domains of PhoQ using disulfide-scanning mutagenesis. Building upon our previous analysis of the periplasmic helix at the dimer interface [35], new single cysteine residue mutations were introduced along the transmembrane helices and at selected positions within the HAMP domain (**Figure 6A** and **B**). Without the oxidizing environment of the periplasm, measuring the extent of disulfide bond

formation in these mutants required the presence of an oxidative catalyst, Cu(II)(1,10-phenanthroline)$_3$ (CuPhen). For each residue in the predicted TM domain, we calculated the fraction of crosslinking from the measured intensities of covalent dimer and monomer bands on a Western blot.

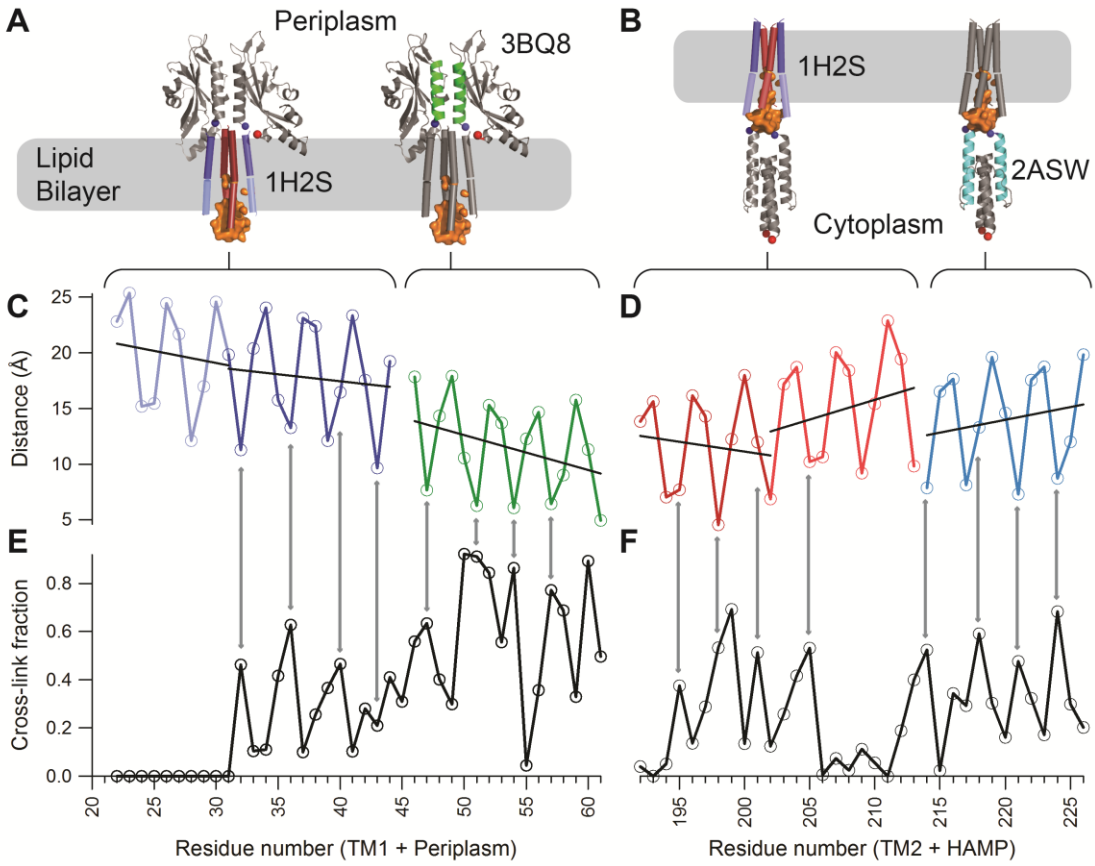

Figure 6 - Comparison of the crosslinking efficiency with structural models. (A) PhoQ TM1-periplasm-TM2 model in a lipid bilayer. The color-coded helical regions (blue-green-red, respectively) indicate where cysteine mutations were made. An orange envelope marks the water hemichannel. (B) PhoQ TM1-TM2-HAMP model in a lipid bilayer. Color-coding (blue-red-cyan, respectively) is applied to the regions probed by cysteine mutations. The water hemichannel is shown as in (A). (C) Inter-monomer distances between the dimeric structures of structural models for TM1-periplasm. The first TM helix is modeled from HtrII (PDB ID: 1H2S) and the periplasmic helix is from *E. coli* PhoQ (PDB ID: 3BQ8). The measured distances are between Cβ-Cβ' of corresponding residues (or Cα-Cα' for glycine). Black lines indicate linear fits

to each helical segment. (D) Inter-monomer distances between dimeric structures of structural models for TM2-HAMP. The second TM helix is from HtrII, and the HAMP helix is from *Archaeoglobus fulgidus* (PDB ID: 2ASW). Distances and fits were done as in (C). (E) Crosslinking data from the full length PhoQ protein in a native membrane for cysteine mutants 22 through 61. (F) Crosslinking data from the full-length PhoQ protein in a native membrane for cysteine mutants 192 to 226.

3.3.1 – Comparison of disulfide crosslinking efficiency to homologous crystal structures

The crosslinking efficiency should depend inversely on the distance between the reacting thiol groups [67], so in an initial modeling approach, we compared the measured crosslinking efficiency for all three domains against their individual structures or homologous structures. To model a full-length PhoQ, we mapped the periplasmic crosslinking data on the crystal structure of the PhoQ periplasmic sensor, the transmembrane crosslinking data on the transmembrane structure of HtrII, and the cytoplasmic crosslinking data on the HAMP structure of Af1503 from *Archaeoglobus fulgidus* (**Figure 5B**). These comparisons test how faithfully these individual domains represent the full-length structure of PhoQ (**Figure 5C**). Importantly, the crosslinking data also adds new structural insight by spanning the intact juxtamembrane regions, which were not present in previous single domain structures.

We compared the inter-residue distances in the transmembrane helical bundles of HtrII with the corresponding experimental crosslinking data (**Figure 6**). The transmembrane four-helix bundle of HtrII and other HKs are oriented with the N-terminus of TM1 and the C-terminus of TM2 directed towards cytoplasm. The core of the HtrII bundle is well packed near the periplasm, but its helices kink and diverge slightly near the cytoplasm. The crosslinking fractions agree qualitatively with this bipartite structure. Near the periplasmic end of the bundle, we observe a periodic pattern of crosslinking efficiency, close to that seen for an ideal α-helix, which repeats

with a period of 3.6 residues. Fitting a sinusoidal function to the data resulted in a period of 3.5 residues for TM2 and 3.7 residues for TM1 (**Table 3**). We computed a phase offset to determine if there was relationship between variation in crosslinking efficiency and the expected distance variation for an alpha helix.

Table 3 – Least-squares fitting of a sinusoidal function to the crosslinking efficiency of PhoQ and the inter-residue distances of PhoQ, HtrII and Af1503 crystal structures.

Helix	Period[1]	Phase offset[2]
PhoQ TM1	3.67 ± 0.13	173°
HtrII TM1	3.69 ± 0.03	
PhoQ TM2	3.53 ± 0.30	168°
HTrII TM2	3.67 ± 0.08	
PhoQ HAMP	3.53 ± 0.20	153°
AF1503 HAMP	3.54 ± 0.02	

[1] Number of residues per repeat

[2] Differences in phase for the fitted sinusoidal waves between the experimental crosslinking data and the inter-monomer distance data (Cβ-Cβ' distance or Cα-Cα' for Gly) taken from corresponding crystal structure

There was little, if any, crosslinking observed in the cytoplasmic end of the bundle along the polar cavity of PhoQ (**Figure 6E** and **F**). Thus, the low degree of crosslinking near the cytoplasmic side of the bundle agrees with the presence of a water hemichannel, shown as solvent accessible surface in **Figure 6A** and **B**. However, the complete lack of crosslinking on the cytoplasmic side of PhoQ TM1 helices cannot be explained by the HtrII structure. The lack of cross-linking suggests a larger separation in the PhoQ hemi-channel compared to that in HtrII. At the periplasmic side of the TM bundle, we observed that the TM1 helices crosslink as strongly as the TM2 helices, despite the TM2 helices being closer together in the HtrII crystal structure.

Taken together, these data indicate that HtrII is only an approximate model for the TM domain of PhoQ.

The structure of the functional dimeric form of the periplasmic sensor domain has previously been determined [17] and validated in full-length PhoQ by disulfide crosslinking [35]. However, this structure is missing the short linker that connects TM1 to the N-terminal helix of the sensor, whereas our new crosslinking data does provide structural information in this region (**Figure 6E**). Interestingly, the crosslinking efficiency maintains a sinusoidal variation with a consistent phase through this linker (**Figure 7A** and **B**), suggesting that TM1 and the N-terminal helix of the sensor domain form a single uninterrupted helix. The similarity of their phases can be appreciated qualitatively by inspection of the data, or quantitatively by fitting the data for TM1 and the periplasmic helix to a sinusoidal function (**Figure 7**). The computed phases for the two structures match within experimental error. However, there is a small deviation at the junction of the two helices near residue 43, which might reflect a slight kink or bend in the helix as it emerges from the membrane.

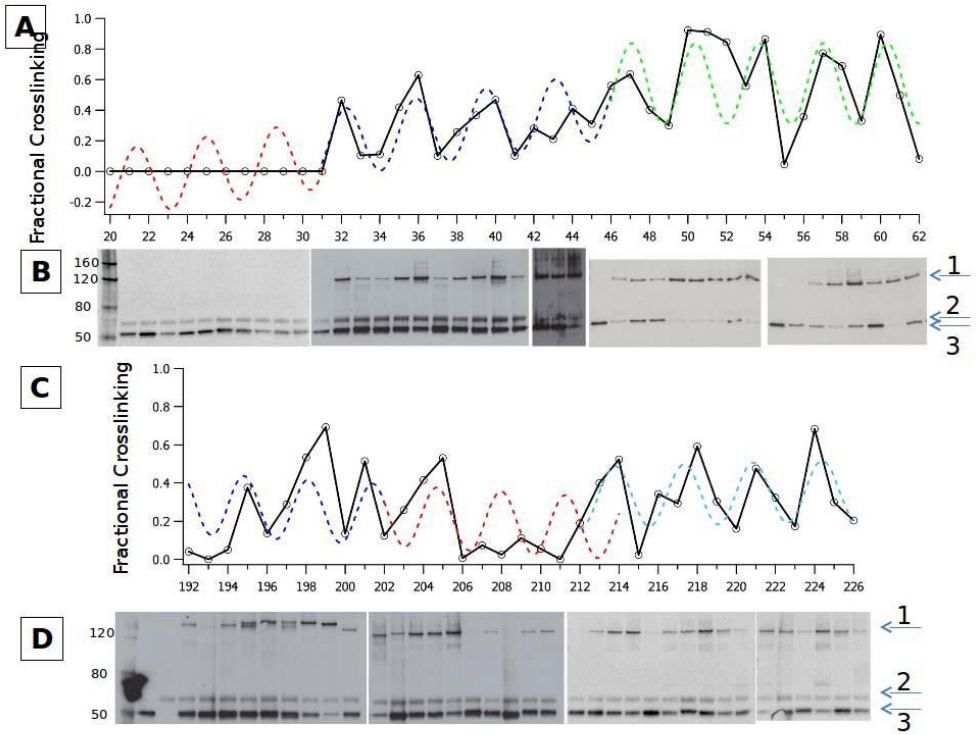

Figure 7 - Analysis of the fractional crosslinking of PhoQ residues. (A) Fractional crosslinking of PhoQ residues 20-62 (black lines with circles) are fitted using a sine wave over the regions that correspond to the domains of PhoQ (dashed lines) the colors are maintained from Figure 2: dark blue is the well-packed domain, red are residues that line the cytoplasmic cavity and green are the HAMP residues. These data demonstrate a right-handed helix (ω=3.62) for TM1 that is in phase with the previously reported data. (B) Representative western blots of PhoQ residues reported in A. Each lane represents the data point directly above it in A. Arrows on the right of the figure indicate 1) the crosslinked PhoQ dimer band 2) an E coli lysate band 3) PhoQ monomer band. (C) Fractional crosslinking of PhoQ residues 192-226 are fitted using a sine wave over the regions that correspond to the domains of PhoQ (dashed lines) the colors are maintained from Figure 2: dark blue is the well-packed domain, red are residues that line the cytoplasmic cavity and light blue is the HAMP. These data demonstrate TM2 is a left-handed helix (ω=3.29) with a striking lack of continuity with the first HAMP helix due to a disturbance in the phase of the helix which arises from residue P208. (D) Representative western blots of PhoQ residues reported in (C). The numbering of the arrows on the right is identical to (B).

Another short linker of unknown structure connects TM2 to the HAMP. This region corresponds

to the cytoplasmic, divergent end of the bundle, resulting in reduced crosslinking over a 6-

residue segment spanning residues 206-211. A conserved Pro residue at position 208 of PhoQ is likely to lead to a bend in the helix in this region [61,104]. The sine waves fitted to the crosslinking data within TM2 are out of phase with respect to the HAMP helix, suggesting the linker between these two regions either adopts a distorted helical or non-helical geometry (**Figure 7C** and **D**).

3.3.2 – Multi-state Bayesian modeling

We collected data on the full-length PhoQ protein in a native membrane, which was free to structurally fluctuate between signal transduction states. Therefore, we do not assume that all crosslinking experiments necessarily probe a single structural state. For example, one structural state cannot explain both high TM1-TM1' crosslinking (residues 32-45) as well as high TM2-TM2' crosslinking (residues 192-206) (**Figure 6E** and **F**) without introducing steric clashes. Consequently, we hypothesize the presence of multiple, distinct structural states in the sample. We used a multi-state Bayesian modeling of cysteine crosslinking data [9], which simultaneously models several structures based on experimental and prior information (such as the available structural information), and infers additional parameters (*e.g.*, uncertainty in the data, σ_0, and population fractions, w_i).

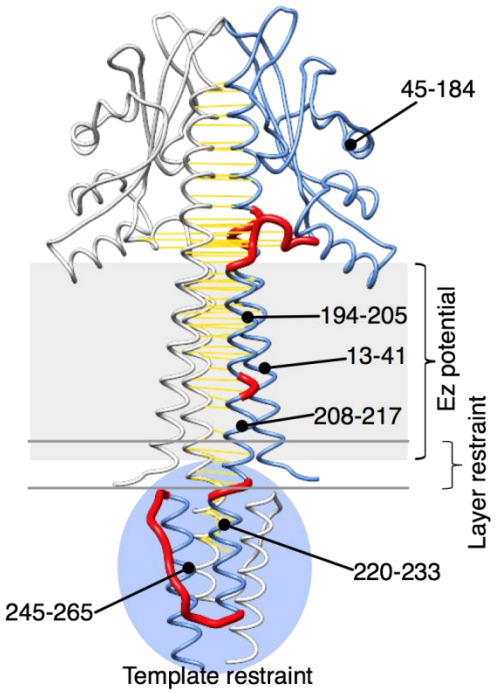

Figure 8 - Representation and score. Each chain of PhoQ homodimer was divided into 6 rigid bodies (represented by a blue trace in one monomer) and 5 short intervening flexible segments (red traces). Rigid body segments are indicated by the corresponding residue ranges. Each residue was represented by a bead centered on its Cα atom. The score terms included the likelihood score for cysteine cross-linking data (yellow lines) as well as for template structure information (applied to the HAMP domain, light blue circle), a statistical potential to enforce the correct stereochemistry of the flexible segments, the V_{Ez} potential to account for the membrane environments (applied to TM region, grey box), the layer restraint to anchor residue F17 to inner leaflet (black lines), and the excluded volume.

We divided the PhoQ dimer into 6 rigid bodies for each monomer, for a total of 12 rigid bodies in the dimer (Materials and Methods and **Figure 8**). A coarse-grained representation of PhoQ was used, in which each residue is modeled as a bead centered on the Cα atom. The conformations of the dimer were explored without imposing any symmetry between the two chains, using a Gibbs sampling scheme relying on a Monte Carlo algorithm enhanced by Replica

Exchange [79]. The sampled models were clustered based on the predicted cross-linked fractions. Thus, members of the same cluster predict similar data, although they might be structurally different, especially in regions that are not restrained by the data (**Figure 9** and **Figure 8**).

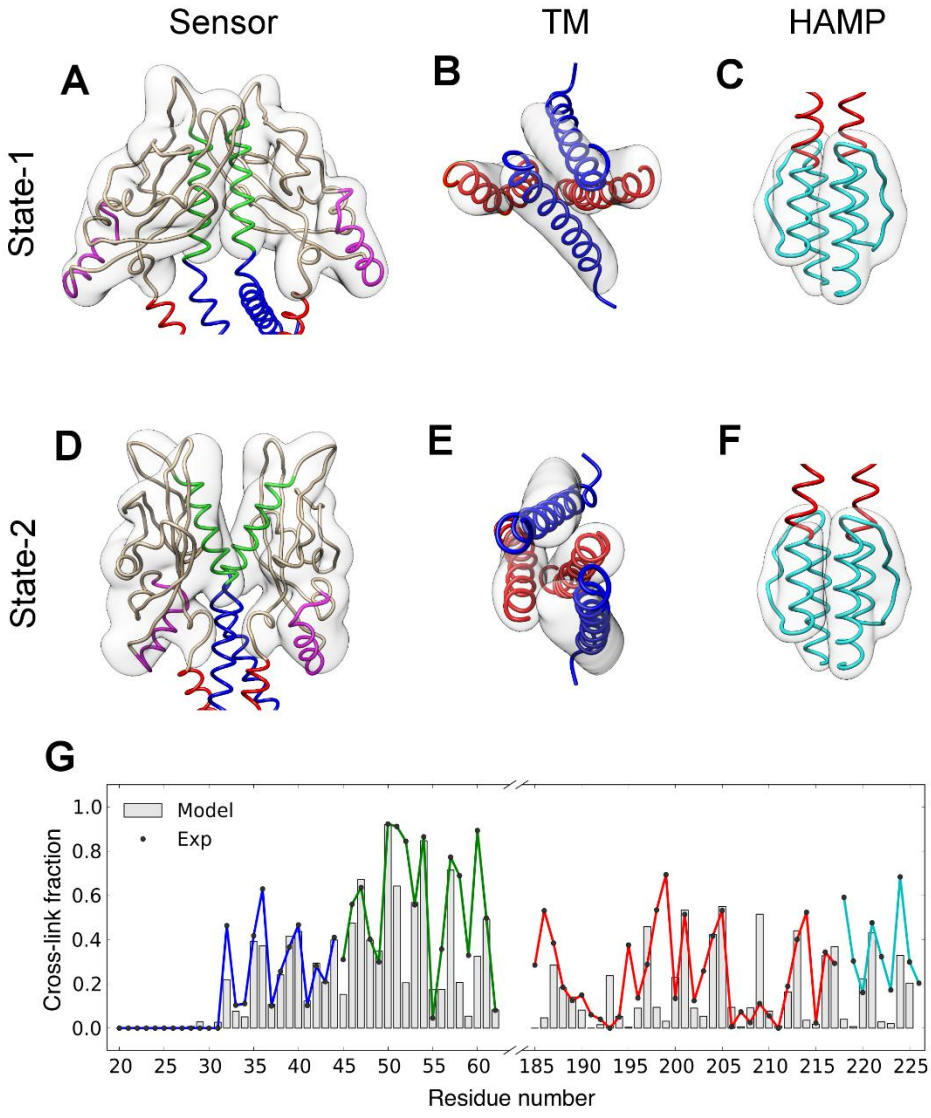

Figure 9 - Analysis of the most populated cluster found in 2-state modeling. Backbone ribbon representation of the cluster representative of: (A) sensor domain in State-1; (B) TM domain in

State-1 as viewed from the sensor domain looking into the cytoplasm; (C) HAMP domain in State-1; (D) sensor domain in State-2; (E) TM domain in State-2, viewed from periplasm looking into the cytoplasm; (F) HAMP domain in State-2. The cluster structural variability is represented by the transparent density volumes calculated using the VMD VolMap tool [46] . The color-coding for (A-F) is as follows: periplasmic sensor helices (residues 45 to 61) are in green, the Mg^{2+}-binding, acidic patch (residues 137 to 150) in magenta; the TM1 and TM2 helices of the TM domain are in blue and red, respectively; and the HAMP domain in cyan. (G) Overlay of model data, predicted by the highest likelihood model of the cluster (grey bars), and experimental cross-linked fractions, color-coded by the domain definition above.

Below, we focus structural analysis on the most populated cluster, which corresponds to the peak with the greatest probability in the posterior probability distribution of states, given the cross-linking data and domain models. Cluster representatives and predicted cross-linked fractions for all clusters with a population greater than 3% are reported in **Figure 8** and **Table 4**.

To predict the minimal number of states that best explain the crosslinking data, the Bayesian approach was applied independently for 1, 2, and 3 states.

Table 4 - Properties of the clusters with population greater than 3% found with 1-state, 2-state and 3-state modeling: cluster population, average and best χ2 and likelihood score (-log p(D|M,I)).

Number of states	Cluster id	Cluster population	Center		Best	
			χ^2	*L*	χ^2	*L*
1	1	0.16	0.85	-33.70	0.75	-46.64
	2	0.06	0.94	-30.47	0.79	-39.69
	3	0.05	0.91	-35.26	0.80	-43.99
	4	0.04	0.78	-43.38	0.66	-57.33
	5	0.03	0.79	-39.72	0.71	-50.21
2	1	0.12	0.65	-59.60	0.54	-67.78
	2	0.10	0.84	-29.91	0.70	-49.31
	3	0.06	0.73	-38.59	0.64	-57.11
	4	0.05	0.83	-31.29	0.69	-44.36
	5	0.04	0.71	-48.21	0.59	-61.72
	6	0.04	0.79	-26.59	0.69	-44.70
	7	0.04	0.77	-33.16	0.67	-44.82
3	1	0.16	0.79	-23.85	0.69	-43.50
	2	0.06	0.64	-48.47	0.55	-62.76
	3	0.06	0.84	-31.42	0.65	-53.75
	4	0.05	0.82	-16.56	0.73	-45.09
	5	0.04	0.76	-36.36	0.65	-51.36
	1	0.16	0.79	-23.85	0.69	-43.50
	2	0.06	0.64	-48.47	0.55	-62.76

3.3.2.1 – 1-state modeling.

The cluster analysis of the sampled models (**Table 4**) revealed that the experimental data could not be fully explained by a single structure. The 1-state model was in good agreement with the predicted cross-linked fractions in the periplasmic side of TM1 (residues 13 to 45) and cytoplasmic side of TM2 (residue 205 to 215). However, the model does not match a large region of data with a high cross-linked fraction, the periplasmic side of TM2 (residues 195 to 205). The reason is that a single structure cannot simultaneously reconcile high crosslinking on the periplasmic side of both TM1 *and* TM2. Instead, for the TM2 periplasmic region, the model-predicted cross-linked fractions equal to zero. Therefore, in the 1-state model, proximity between the periplasmic region of TM2 and TM2' is not observed due to steric exclusion by the TM1 and TM1' helices.

3.3.2.1 – 2-state modeling

The most populated cluster of two states found by 2-state modeling explained crosslinking data better than 1-state modeling, as shown by the lower likelihood score (**Table 4**) and the improved agreement between the model and the data for the periplasmic TM2 region and surrounding residues (185-205) (**Figure 9**). The inferred population fractions of State-1 and State-2 were 40.5% and 59.5%, respectively. The two states differ at the dimeric interface in the arrangement of the helices from every domain.

For the periplasmic region, State-2 resembles the crystallographic structure of the PhoQ sensor domain (Cα RMSD = 3 Å), 3BQ8, previously proposed to correspond to the activated state [17]. In contrast, in State-1 the periplasmic helices are closer to a parallel configuration. The

periplasmic helices transition between a parallel (State-1) and a crossing configuration (State-2); this transition corresponds to a scissoring motion. A consequence of the scissoring motion is a displacement of the acidic patch (residues 145-154) in the periplasmic domain, from resting on the surface of the membrane in State-1 to a position deeper in the membrane in State-2.

The scissoring motion of the periplasmic, interfacial dimer helices propagates into the TM domain. This motion is best seen on the periplasmic side of the TM bundle (top down view of helical bundle in **Figure 9B** and **E**), where the pairs of helices take turns displacing each other. State-1 predicts that the TM1 and TM1' helices (blue) pack close and displace the TM2 and TM2' helices (red), while in State-2 the TM2 and TM2' helices move towards the center of the bundle and displace the TM1-TM1' intersubunit helical contacts. This displays how the scissoring motion propagates within the bundle, because the scissoring towards the bundle center of one helix pair causes the scissoring out of the other pair. The 2-state modeling explains the high crosslinking observed in the same region by postulating the existence of a mixture of states.

For the TM domain in particular, the modeling added valuable insight. The cytoplasmic regions of both TM1 and TM2 show low crosslinking (thought to be due to a water pocket, colored orange in **Figure 6A** and **B**) and therefore were not as structurally constrained as regions that show more regular periodic crosslinking (periplasmic and HAMP helices).

The large changes seen in the TM domains are coupled with smaller changes in the HAMP domains. Specifically, the scissoring displacement seen in the TM domain is also observed for the HAMP helices. In State-1, the helix 1-helix 1' distance is shorter than the helix 2-helix 2' distance near the N-terminal end of the bundle; this relationship reverses in State-2 (**Figure 9C**

and **F**). Presumably, this conformational change is coupled to additional, previously characterized changes in the catalytic and DHp domains [2,27].

3.3.2.3 – 3-state modeling

Models in the most populated clusters in 3-state modeling explain the data worse than those found by 1-state and 2-state modeling, as indicated by the average and best likelihood scores for the clusters (**Table 4**). The previously identified models were not found here because we imposed a lower bound of 0.2 on the individual population fractions w_i (Materials and Methods).

3.3.2.4 – Selecting the best model

A single state model does not explain all the cysteine cross-link fractions, thus strengthening the hypothesis that the sample contains multiple conformations of PhoQ. The 2-state model fits the data significantly better than either 1- or 3-state models. The 2-state model suggests a reorientation of the periplasmic domain accompanied by a propagated scissoring movement through the TM and HAMP. Therefore, the 2-state model is the most parsimonious explanation of the data.

3.3.2.5 – Deviations between crosslinking data and the 2-state model

While the 2-state model best fits the experimental observations, a few data points still could not be explained. In particular, isolated deviations were observed at residues 52, 195, 199, 208, 209, and 218 (**Figure 9G**). These discrepancies can in principle originate from inaccuracies of the Bayesian model (including the forward model, noise model, sampling, and the assumed number of different states) or the representation of the system. To discriminate between these

possibilities, we investigated the phenotypes of the cysteine mutants, by measuring transcriptional activity at low and high Mg^{2+} concentration (**Figure 10**), as described previously [68].

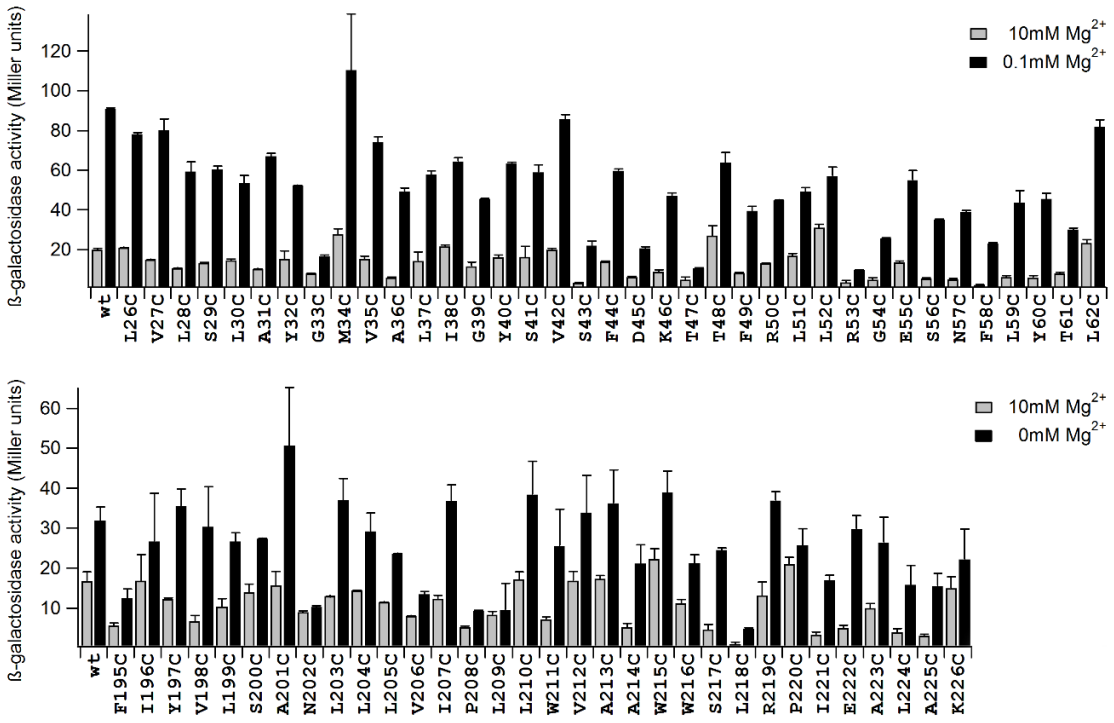

Figure 10 Phenotypic changes in response to Cys mutations in PhoQ. We assessed the activity of Cys mutants by the Miller assay [68]. TIM206 (*mgtA::lacZ* Δ*phoQ*) cells were transformed with a plasmid encoding PhoQ and a Cys mutation at a single position. Mutation to Cys was tolerated at most positions. Positions known to have a critical function also have no activity when mutated (*e.g.,* 202). Many positions where the Bayesian model does not explain the data (residues 195,208, 209, and 218) also do not respond to Mg^{2+}.

The mutants P208C and L209C have low β-galactosidase activity at both high and low concentrations of Mg^{2+}. By contrast, the wild-type protein activity changes 2-5 fold between these Mg^{2+} concentrations. Interestingly, a kink in TM2 occurs between P208 and L209 in an MD model of the TM domain of PhoQ [61], suggesting this region is a fulcrum of movement. For

these positions, we hypothesize that these cysteine mutations abolish signal transduction because they tamper with the helical kink.

Similarly, F195C and L218C showed low activity at both Mg^{2+} concentrations. Both mutants are positioned to break the chain of signaling because they lie in transition areas between domains: residue F195 is between the periplasm and the TM, while L218 is in the loop between the TM and HAMP. For F195C, the helical period between experimental and model data is shifted, which indicates a potential helical rotation in that region (residues 195 to 199). This portion of TM2 was modeled as a rigid body extending from 194-205, but the discrepancy suggests that two rigid bodies or a flexible chain might be more appropriate representations for this region. For L218C, this mutation was part of the loop region unconstrained by existing crystal structure data, which could account for why the model did not accurately predict its crosslinking.

Residue 52, on the other hand, shows activity similar to wild-type at low Mg^{2+}, but does not agree with the 2-state model. Because this mutant has reasonable activity data, we focused on the crosslinking data and observed an unusually broad peak of high crosslinking for residues 50-52, which is inconsistent with the helical period seen in the crystal structure (PDB ID: 3BQ8). This discrepancy encouraged us to repeat the previously published crosslinking experiments for a portion of the periplasmic helix at the dimer interface. In this region, disulfide crosslinks occur spontaneously and do not require the aid of an oxidant like CuPhen, as is required for HAMP and TM domains. The previous periplasmic crosslinking experiments [35] used long, overnight incubations in LB medium. However, when we incubated for shorter periods of time (to avoid spurious crosslinks) in minimal medium (for precise control of Mg^{2+} concentration), we found

that residue 50 has not crosslinked nearly as much as residue 51 and that the extent of crosslinking of residue 52 was dependent on the Mg^{2+} concentration (**Figure 11**). The reduced crosslinking for both residues improves agreement with an ideal helical period (**Figure 11**, black dotted line). This indicates that the Bayesian analysis had identified an artifact in the data, corresponding to a cross-link that stabilizes a non-native conformation occurring only after a long incubation time.

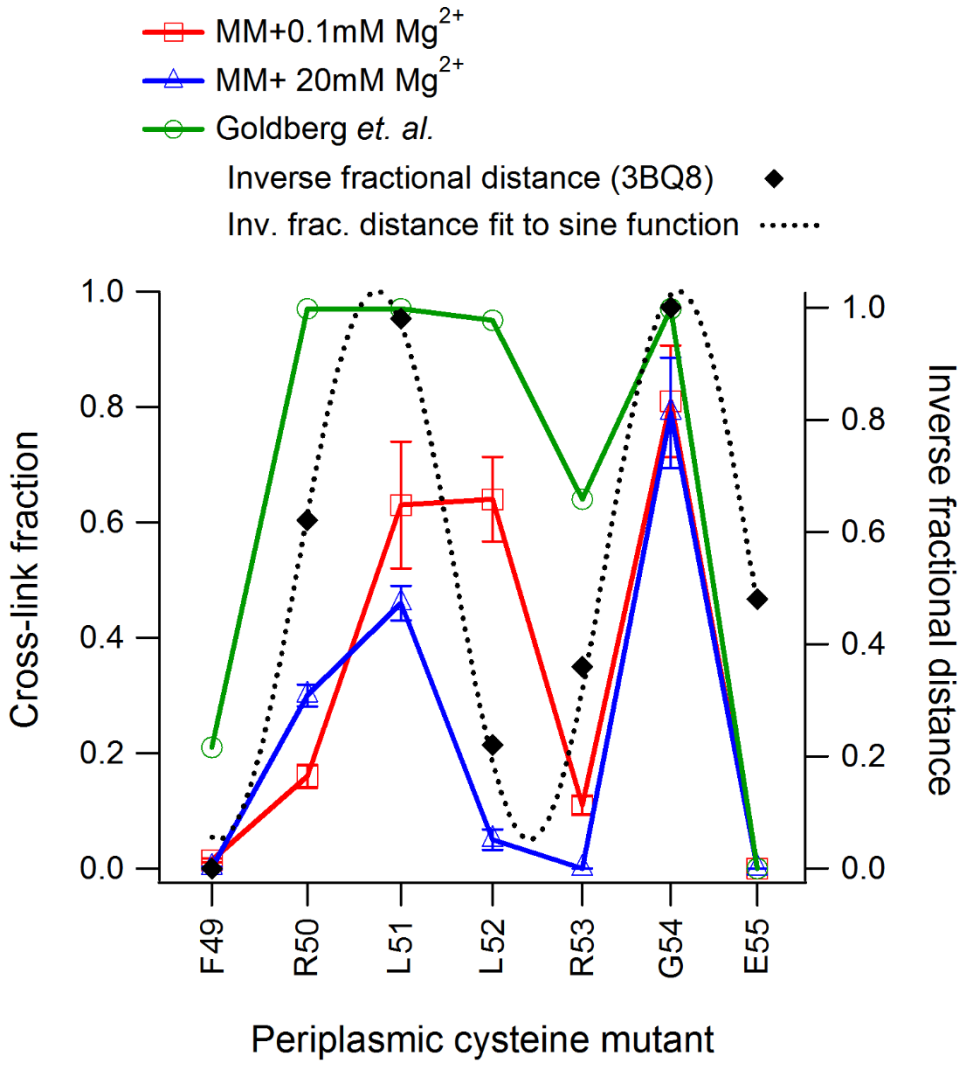

Figure 11 - Comparison of crosslinking efficiency for the periplasmic helix under different conditions. (Left axis) Green curve is original crosslinking data **[35]**. Blue and red curves are new data collected in minimal media at mid-log growth. Error bars represent the standard deviation of triplicate experiments. (Right axis) Black dotted line is a sinusoid fit to the inverse, fractional inter-monomer distances as measured between Cβ-Cβ' (or Cα-Cα' for Gly) of 3BQ8, represented by black diamonds.

In summary, Bayesian modeling helped us rationalize flaws originating from artifactual disulfide formation (residue 52), inactive constructs (residues 208 and 209), representation inaccuracies (residue 195), and loop regions not sufficiently constrained by structural data (residue 218). In

traditional modeling, these points would be considered as outliers and removed from the data set. In the Bayesian framework, such a manual intervention is not necessary because an uncertainty parameter is associated to each data point, thus allowing those points that are not consistent with the bulk of the data to be properly down-weighted in the construction of the model. Instead, the 2-state model motivated additional functional experiments to explain the large differences between the observed and predicted data.

3.3.3 – Structural variation between signaling states

The 2-state model proposes conformational changes between State-1 and State-2 larger in amplitude and more closely related to a scissoring motion than the anticipated motions from the piston and gearbox models [16,27]. To test whether or not the scissoring motion is unique to PhoQ, we quantified the structural variability among the known structures of two-component HK domains. A number of crystallographic and NMR structures of dimeric HK domains are available, and we selected the 4-helix bundles where each dimer contributes two helices. We also required that domains had multiple structures in multiple conformations, to be able to infer movement from the crystal structures (**Table 5**). First, we examined the structures used to propose the original piston shift and gearbox motion models. The piston shift was originally described based on the aspartate sensor, Tar [16], while the gearbox model was based on HAMP structures including a HAMP(Af1503)/DHp(EnvZ) chimera [27]. Next, we also quantified the structural variability seen in the citrate sensor domain in the presence and absence of citrate [49], as well as the DesK DHp structures believed to represent different signaling states [2]. To compare the structural variability in these two-component domains, we identified pairs of

symmetry-related helices likely to be relevant to signal transduction, either as judged by the original authors [16,27] or by virtue of the helices connecting adjacent domains.

We describe the changes in helix orientation by relying on six independent degrees of freedom that define a convenient coordinate system. Two degrees of freedom match previous signaling models; a translational motion parallel to the bundle axis ("height"=z) corresponds to the piston model, and a rotation about the helix axis ("helix phase"$= \psi$) corresponds to the gearbox model. The remaining four degrees of freedom are helix tilt towards the bundle axis ("towards tilt"=φ_1), helix tilt perpendicular to the "towards tilt" ("sideways tilt"=φ_2), radial displacement from the central bundle axis ("radius"=r), and global rotation of individual helices relative to their neighbors around the central bundle axis ("bundle phase"=θ) (**Figure 12**). We classify radial displacements (r) and towards tilt (φ_1) as scissoring motions.

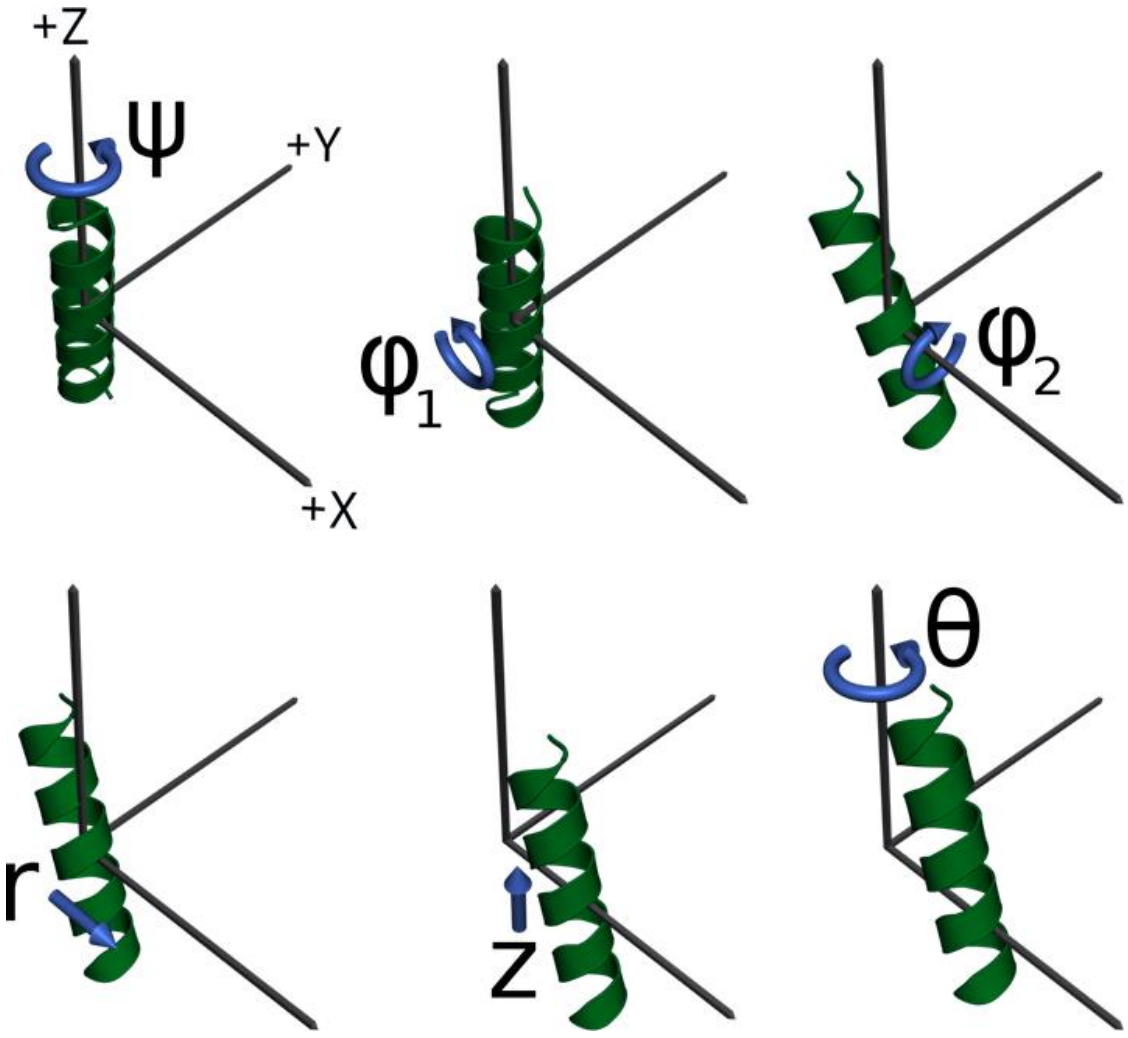

Figure 12 - The six degrees of motion in the order they are applied to fit any given helix: ψ: first rotation about Z axis; ϕ1: rotation about Y axis; ϕ2: rotation about X axis; r: translation along X axis; z: translation along Z axis; θ: second rotation about Z axis.

To compare contributions of each change, we measured the variation along each degree of freedom and normalized all measurements to displacements. For every pair of helices, we defined the two largest changes contributing to the total RMSD variation in **Table 5**. For each of the displacements in the table, two distances are given (*e.g.* Tar Sensor largest motion, φ_2: 2.5/2.3 Å). Those distances correspond to a measurement of change in each of the dimers that

make up the 4-helix bundle (*e.g.*, the largest difference between two symmetry-matched helices in chain A was 2.5 Å, and 2.3 Å refers to the same in chain B).

Table 5. The largest quantified changes between pairs of correlated helices in two-component domains.

Domain	Chains/ Residues	Largest change (Å)	2nd largest change (Å)	PDB IDs
Tar Sensor	(A,B)/ (42-57)	sideways tilt (φ_2) 2.5 / 2.3	height (z) 1.8 / 1.6	1MJW, 1LIH, 1VLS, 1VLT, 1WAS, 1WAT, 2ASR, 2LIG
Citrate Sensor (CitA*)	B/ (12-25, 45-51)	radius (r) 5.5 / 5.2	bundle phase (θ) 3.8 / 4.4	1P0Z, 2J80
AF1503 HAMP helix 1	(A,B)/ (280-297)	towards tilt (φ_1) 2.1 / 1.7	helix phase (ψ) 1.7 / 1.8	2LFR, 2LFS, 3ZRV, 3ZRW, 3ZRX
AF1503 HAMP helix 2	(A,B)/ (310-328)	towards tilt (φ_1) 2.1 / 1.4	radius (r) 1.7 / 1.4	same as above
EnvZ DHp	(A,B)/ (333-345)	sideways tilt (φ_2) 3.1 / 2.4	towards tilt (φ_1) 1.6 / 2.1	same as above
DesK DHp	(A,B)/ (182-198)	helix phase (ψ) 3.2 / 3.2	radius (r) 3.0 / 2.7	3EHF, 3EHH, 3EHJ, 3GIE, 3GIF, 3GIG
DesK DHp	(A, B)/ (224-238)	radius (r) 2.4 / 2.9	towards tilt (φ_1) 1.5 / 2.4	same as above

Piston (z) and gearbox (ψ) degrees of freedom are shown in red to illustrate that these are not representative of TCS signal transduction in general.

* Both helices from chain B were measured because all of the displacement was limited to that chain.

The quantified changes for the aspartate sensor domain agree well with the qualitative observations of a "swinging piston" [16] or torque [22] mechanism of signal transduction. For

the Tar sensor helix in both chains A and B, sideways tilt (φ_2) contributes the greatest displacement, followed by piston shift (z).

However, the HAMP(Af1503)/DHp(EnvZ) chimera is more complex than suggested by the proposed gearbox model. For helix 1 of chains A and B of the HAMP domain, helix phase changes (ψ) feature prominently, agreeing with the gearbox mechanism. However, swinging toward the bundle axis (φ_1) contributes even larger changes. In contrast, helix 2 of chains A and B features little helical rotation (ψ) and instead primarily rotates away from the bundle axis (φ_1) and also translates outward (r). We observe a similarly pronounced sideways swinging rotation (φ_2) in chains A and B of the first helix of EnvZ DHp domain, which outweighs other changes, including helix phase changes (ψ). These data indicate that gearbox rotations do not fully explain the structural variation between signaling states.

Moreover, similar changes were observed for the DesK DHp domains. Lateral translations consistently dominate, with the exception of a single pair of helices in DesK, where gearbox rotations dominate and lateral translations come in a close second. Even larger lateral translations appear in CitA, although one of the structures in the analysis (PDB ID: 1P0Z) may not be truly representative of the native dimeric state.

In six out of the seven domains we studied, one of the two largest changes is either a radial displacement (r) or towards tilt (φ_1), both indicative of scissoring, with infrequent contributions from a piston translation or a gearbox rotation. However, this does not necessarily rule out the possibility that the relatively small gearbox and/or piston shift motions might propel the larger changes in other degrees of freedom.

3.4 – Discussion

We present a model of HK signal transduction through the membrane, the first to utilize structural data taken from the full-length, dimeric protein in a native membrane. We generated this model using disulfide scanning mutagenesis and existing homologous crystal structures. Disulfide scanning mutagenesis allowed us to study the full-length protein in its native membrane environment, without relying on isolated domains in micelles or other membrane mimetics. We found good agreement between the crosslinking data and existing structures of water-soluble HAMP, TM, and periplasmic domains, indicating that the isolated domain structures are good models for the corresponding domains within the full-length protein in a native membrane environment as well as that the cross-linking data is accurate. The crosslinking data are almost 180° out of phase with distances derived from homologous crystal structures (**Table 3**), which is expected because crosslinking efficiency decreases over greater distances.

The crosslinking data covers the juxtamembrane regions connecting the TM domain to the sensor and HAMP domains (**Figure 6**). These data provide the first evidence for an uninterrupted helix spanning TM1 to the N-terminal helix of the sensor domain. Additionally, the crosslinking data spanning the TM2-HAMP boundary indicates a possible interruption, which may be either a kinked helix or a disordered linker connecting the two domains. This interrupted structure may be necessary to form the previously described water hemichannel on the cytoplasmic face of the TM 4-helix bundle [34].

Bayesian modeling revealed that the crosslinking experiments likely probed two structural states. We in fact anticipated at least two states for the following two reasons. First, PhoQ

must respond to its environment by relying on a thermodynamic equilibrium between its two signaling states, a prediction that is consistent with the two modeled states predicted to be present in the sample in similar proportions (40.5% State-1 and 59.5% State-2). Experimental data also supports signaling states near equilibrium, where we see a degree of activation of only 2-5 fold in low Mg^{2+} concentrations. These results are in agreement with the EPR studies of Trg from *E. coli*, which identified a dynamic and loosely packed transmembrane domain [6]. Second, only two structural states could explain conflicting crosslinking data within the TM domain, where TM1-TM1' crosslinks are sterically inconsistent with TM2-TM2' crosslinks.

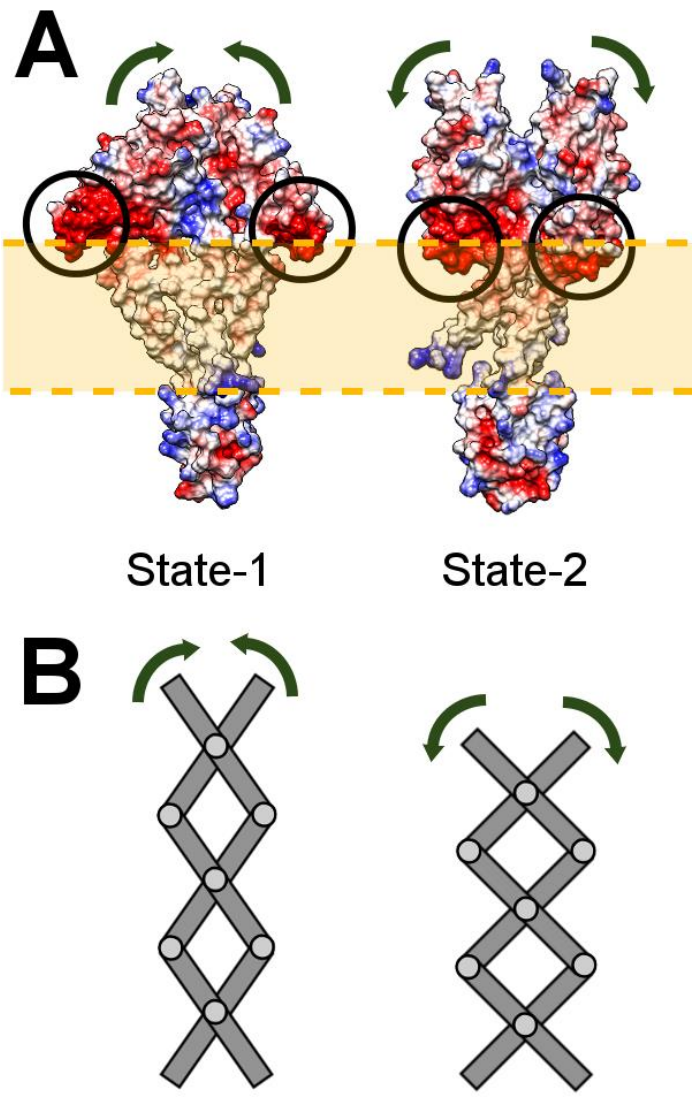

Figure 13 - Cation-binding, acidic patch movements predicted by the Bayesian multi-state modeling. (A) Electrostatic surface representation of the two states of the acidic patch as it moves out of (State-1) and in to (State-2) the membrane bilayer. Surface made with UCSF Chimera **[75]**. (B) Schematic "scissor lift" mechanism of signal propagation.

These two alternative conformations suggest large displacements of the sensor domains that insert or remove the divalent cation-binding acidic patch within the membrane (**Figure 13A**). This change is coupled to scissoring transitions in the sensor, transmembrane, and HAMP

domains, resembling a “scissor lift” mechanism of signal propagation (**Figure 13B**). The modeling predicts large conformational rearrangements in the transmembrane domains, resembling a lateral scissoring change (**Figure 9B** and **E**), where two opposing helices move inward and displace the other two opposing helices which move outward, a pattern which we summarize in **Figure 14A**. We observed similar scissoring motions across several two-component systems (**Figure 14B-D**). This universal scissoring between distinct conformations was seen in the sensor, HAMP, and DHp domains; furthermore, these motions are consistent with the torque motion proposed recently for the blue-light sensing HK, YF1 [22]. Even the HAMP domain, in which the gearbox model was discovered, exhibits the same large lateral changes, as mentioned briefly in the original publication [24].

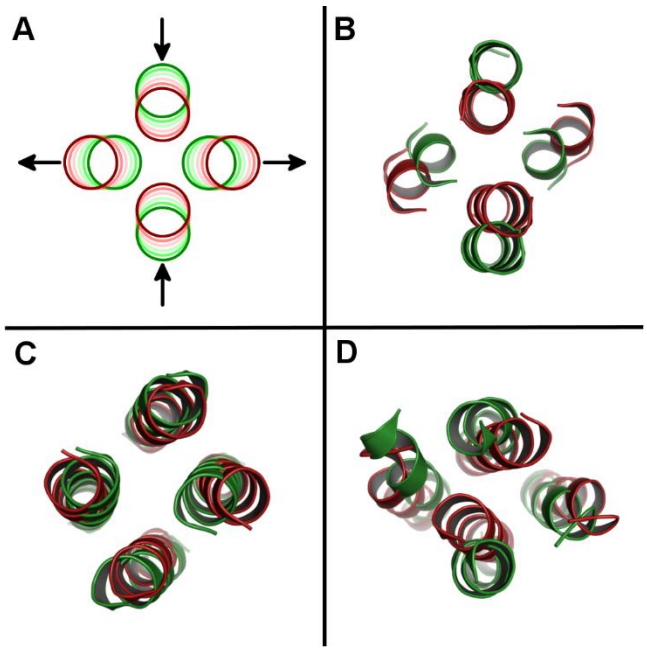

Figure 14 - Scissoring motions across several two-component domains. (A) A helix bundle scissors if two opposing helices move inward and the other two opposing helices simultaneously move outwards. (B) Citrate sensor domain, residues 12-25 and 45-51 (Green: 1P0Z, Red: 2J80).

(C) HAMP domain from AF1503-EnvZ chimera, residues 283-297 (Green: 2L7H, Red: 2Y21). (D) DesK DHp domain, residues 182-198 and 224-238 (Green: 3GIG, Red: 3EHH).

We place these qualitative observations on a more quantitative footing by measuring the variation between pairs of structures along six orthogonal degrees of freedom representing: 1) gearbox rotation about the helix axis; 2) piston shifts that vertically displace helices; 3 & 4) tilting towards and perpendicular to the bundle axis; 5) radial displacement of the helix from the bundle axis; and 6) rotation of the individual helices relative to the others about the bundle axis. In every examined case, we find that these domains, including the 2-state model, are not purely described by one pure motion, yet the tilting and radial displacements are a dominant change in almost every two-component domain that we analyzed (**Table 5**).

We conclude that while the piston shift and gearbox rotations do contribute to conformational adjustment, they do not provide a complete picture of signal transduction. A large, lateral scissoring mechanism plays a critical role in PhoQ signaling, exists in many other TCS signaling domains, and may be a universal way of connecting domains and transmitting signals in these systems.

3.5 – Materials and Methods

3.5.1 – Plasmids

pTrcTIMPhoQHIS was a gift from the lab of Dr. Mark Goulian at the University of Pennsylvania. pACYCPhoQa-3DHIS was created as described previously [34].

3.5.2 – Cell propagation

For crosslinking reactions in the transmembrane and HAMP domains, cells were grown on LB agar or in LB medium at 37°C. For the periplasmic mutants, cells were grown in MOPS minimal medium [72] at 37°C.

3.5.3 – Envelope preparations

A fresh streak of cells was plated onto an LB plate containing 100µg/mL ampicillin. 16 hours later, colonies were picked by sterile loop and used to inoculate 5mL LB + 100µg/mL ampicillin. Cultures were grown at 37°C for 24 hours with vigorous shaking (220 rpm) and pelleted by centrifugation at 3700 x g for 10 minutes at 4°C. Cells were washed by resuspension in 30mM Tris, pH 8 and pelleted as above. Afterwards, cells were treated with 20% sucrose in 30mM Tris, pH 8 for osmotic shock and 10mg/mL lysozyme to remove the cell wall. After 30 minutes incubation at 4°C, the cell envelopes were resuspended in 3mL 3mM EDTA, pH 8 and sonicated briefly to disrupt membranes. The samples were spun at 16000 x g for 30 minutes at 4°C to pellet membranes. The membrane fraction was resuspended in 200µL of 2mM Tris, pH 7.5 and stored for use at -80°C.

3.5.4 – Crosslinking reactions

To crosslink cysteine residues in the transmembrane and HAMP domain, it is necessary to provide an oxidizing reagent. The oxidative catalyst, Cu(II)(1,10-phenanthroline)$_3$ (CuPhen) is a small, membrane-permeable reagent that efficiently catalyzes disulfide bond formation in the membrane [63]. We combined a 10µL sample of cell envelopes with 10µL of buffer containing 2mM or 0.2mM CuPhen for a final concentration of either 1mM or 0.1mM. Reactions were

allowed to proceed for 30 minutes at room temperature. Reactions were stopped by the addition of 20mM N-Ethyl Maleimide (NEM) and 20mM EDTA, and reactions were spun at 16,000 x g at 4°C to concentrate membranes. For the periplasmic domain mutants, we used the natural oxidizing environment of the periplasm to promote disulfide bond formation.

3.5.5 – Western blotting and analysis

Oxidized membranes were reconstituted in 20µL of loading buffer (Invitrogen LDS buffer with 8M urea and 0.5M NEM) and heated for 10 minutes at 70°C. 5µL of sample were loaded onto either a 7% or 3-8% gradient Tris Acetate gel (NuPage®, Invitrogen). Proteins were separated by electrophoresis and were dry transferred to a nitrocellulose membrane (iBlot®, Invitrogen). For crosslinking reactions in the transmembrane region, membranes were washed with TBST buffer (10mM Tris, pH 7.5, 2.5mM EDTA, 50mM NaCl, 0.1% Tween 20) and blocked with 3% BSA in TBST. PhoQ was probed using a penta-His antibody (Qiagen). The His antibody was probed with HRP-conjugated sheep anti-mouse IGg (Pierce). Proteins were visualized by exposure to ECL reagent (Amersham, GE health sciences) for 1 minute and exposure to film for 30-60 seconds. For crosslinking reactions in the periplasmic region, membranes where blocked with TBST and 1% BSA (SNAP i.d.®, Millipore), then probed with penta-His HRP conjugate (Qiagen). Pixel density histograms were generated using the ImageJ software, freely available from the NIH [1], and crosslinking efficiency was determined using the ratio of crosslinked dimer to total visible protein (dimer/(dimer+monomer)).

3.5.6 – Sequence-structure threading and model manipulation

To generate electrostatic maps, we threaded PhoQ's sidechains on to our 2-state models using Scwrl [12] and minimized the side-chain using Rosetta fast-relax [51]. Structure visualization and manipulation was performed using PyMol molecular viewer (Schrodinger).

3.5.7 – Multi-State Bayesian Modeling

The modeling and analysis were carried out with the open source *Integrative Modeling Platform* package (IMP; http://www.integrativemodeling.org) [3,83]. IMP can construct structural models of macromolecular protein complexes by satisfaction of spatial restraints from a variety of experimental data .

3.5.7.1 – Representation of the system and initial model

We generated a Cα model of the PhoQ dimer by assembling the models of HAMP, TM, and periplasmic domain. A comparative model of the HAMP domain dimer was created by using as a template the dimeric HAMP-DHp fusion A291V mutant (PDB ID: 3ZRW). A comparative model of the TM monomer was built by using as a template the two helices in the crystal structure of HtrII (PDB ID: 1H2S), corresponding to residues 23-82 of chain B, as a template. The model of the TM dimer was then obtained by applying the crystallographic C_2 symmetry about the dimer axis, observed in 1H2S. The dimer models of the three domains were positioned relative to each other into an initial dimer model of the whole PhoQ using UCSF Chimera [75], subject to the polypeptide chain connectivity between the three domains in each monomer (**Figure 8**). For the subsequent sampling, each monomer was decomposed into 6 rigid bodies and 5 short intervening flexible segments. Rigid bodies included the following segments: 13-41 (TM1), 45-184 (periplasmic rigid body), 194-205 (N-terminus of TM2), 208-217 (C-terminus of TM2), 220-233 (N-terminal HAMP domain rigid body), and 245-265 (C-terminal HAMP domain rigid body). TM2 was divided into two rigid bodies due to a potential kink at P208. The two chains of the PhoQ dimer were sampled without enforcing any symmetry.

3.5.7.2 – Bayesian model of cysteine crosslink data

The Bayesian approach [40] estimates the probability of a model, given information available about the system, including both prior knowledge and newly acquired experimental data. When modeling multiple structural states of a macromolecular system, the model M includes a set X of *N* modeled structures $\{X_i\}$, their population fractions in the sample $\{w_i\}$, and the additional parameters $\{\alpha_n\}$ defined below. Using Bayes theorem, the posterior probability $p(M|D,I)$ of model M, given data D and prior knowledge I, is

$$p(M|D,I) \propto p(D|M,I) \cdot p(M|I)$$

where the likelihood function $p(D|M,I)$ is the probability of observing data D, given M and I; and the prior $p(M|I)$ is the probability of model M, given I. To define the likelihood function, one needs a forward model $f(X)$ that predicts the data point that would have been observed for structure(s) X, and a noise model that specifies the distribution of the deviation between the observed and predicted data points. The Bayesian and likelihood scores are the negative logarithm of $p(D|M,I) \cdot p(M|I)$ and $p(D|M,I)$, respectively.

3.5.7.2.1 – Forward model

The forward model [9] predicts the cross-linked fraction of cysteine pair n after a reaction time t, for a mixture of *N* states $\{X_i\}$:

$$f_n(\{X_i, w_i\}) = \sum_{i=1}^{N} w_i \left(1 - e^{-\alpha_n \rho(r_n)}\right)$$

where $\alpha_n = k_n\ t$ is the product of the unknown intrinsic reaction rate of cysteine pair n and the total reaction time. $\rho(r_n)$ is an efficiency term that depends on the distance r_n between the cysteine Cα atoms and it is computed by considering (i) the uncertainty in the position of the residue centroids along the main chain due to the limited precision in determining the position of the residues, (ii) the cost of having a disulfide bond geometry far from the ideal one, and (iii) the reduction of the reaction volume due to the presence of proximal components and moieties.

3.5.7.2.2 – Likelihood function

The likelihood function $p(D|M,I)$ for dataset $D = \{d_n\}$ of N_{XL} independently measured cross-linked fractions is a product of likelihood functions for each data point. Because the cross-linked fractions vary between 0 and 1, we modeled the noise with a normal distribution truncated to this interval. The likelihood for data point d_n can thus be written as:

$$p(d_n|\{X_i, w_i\}, \alpha_n, \sigma_n) = Z^{-1}\exp\left(-\frac{[d_n - f_n(\{X_i, w_i\})]^2}{2\sigma_n^2}\right)$$

where the uncertainty σ_n shapes the likelihood function and Z is the normalization factor. To account for varying levels of noise in the data, each data point has an individual σ_n.

Furthermore, to encode template structure information for the HAMP dimer domain (residues 235-263), a likelihood function with log-normal noise was defined based on the distances r_{jk} between all Cα atoms that are below 8 Å in the template (PDB code 2Y20):

$$p(r_{jk}|\{X_i\},\sigma_H) = \prod_i Z^{-1}\exp\left(-\frac{\log^2 r_{jk}/r_{jk,i}}{2\sigma_H^2}\right),$$

where $r_{jk,i}$ is the distance between atom j and k in the modeled structure X_i and σ_H is the uncertainty.

3.5.7.2.3 – Prior Information

The prior on a structure is defined as $p(\{X_i\}) \propto \exp(-\sum_i V(X_i))$ where V is a sum of spatial restraints: $V = V_{excl.\ vol.} + V_{C_\alpha\ bonds} + V_{C_\alpha\ angles} + V_{C_\alpha\ dihedrals} + V_{Ez} + V_{layer}$.

The excluded volume restraint $V_{excl.vol.}$ was implemented as a pairwise hard-sphere repulsive potential, where the volume of each Cα particle equals the volume of the corresponding amino acid residue [76]. The bond, angle, and dihedral terms $V_{C\alpha\ bonds}$, $V_{C\alpha\ angles}$, and $V_{C\alpha\ dihedrals}$, respectively, are statistical potentials that enforce the correct stereochemistry, as well as the correct secondary structure propensity, of the flexible backbone [see Supplementary Information]. The V_{Ez} potential [85] was used to model the membrane environment. Furthermore, residues F17 of the two PhoQ chains were confined inside a layer representing the inner leaflet of the membrane, by using a flat bottom harmonic restraint acting on the z coordinate between -17 Å and -13 Å, V_{layer}.

Crosslinking data was collected in three separate experiments for the periplasmic, membrane, and cytoplasmic domains. We used three α_n parameters to model experimental variation between these three data subsets. The priors for α_n are bounded uniform distributions: the lower bound was determined by the highest observed fraction in the subset and the upper

bound by the highest detectable fraction. The priors for σ_n are unimodal distributions [90]: $p(\sigma_n|\sigma_0) = \frac{2\sigma_0}{\sqrt{\pi}\sigma_n^2}\exp\left(-\frac{\sigma_0^2}{\sigma_n^2}\right)$, where σ_0 is an unknown experimental uncertainty; the heavy tail of the distribution allows for outliers. The priors for w_i were uniform distributions over the range from 0 to 1, with the constraint $\sum_i w_i = 1$. Furthermore, a lower bound at 0.2 was enforced on each w_i to avoid visiting conformations already sampled at smaller N values. A Jeffrey's prior $p(\sigma_H) = 1/\sigma_H$ was used for the uncertainty parameter of the likelihood used to incorporate template structure information.

3.5.7.3 – Sampling

A Gibbs sampling scheme based on Metropolis Monte Carlo [79] enhanced by replica exchange was used to generate a sample of coordinates $\{X_i\}$ as well as parameters α_n and w_i from the posterior distribution of a given number of structures (*N*). The moves for $\{X_i\}$ included random translation and rotation of rigid parts (0.15 Å and 0.03 radian maximum, respectively), random translation of individual beads in the flexible segments (0.15 Å maximum), as well as normal perturbation of the parameters α_n and w_i. To facilitate the sampling of the posterior probability, we eliminated its dependence on the uncertainties σ_n by numerical marginalization (Sivia and Skilling, 2006).

Analysis. The set of sampled models $\{M_j\}$ were clustered [21] based on the value of the forward model $f_n(M)$, using the following data-based metric:

$$\left|\left|M_1 - M_2\right|\right|^2 = \frac{1}{N_{XL}}\sum_{n=1}^{N_{XL}}\frac{[f_n(M_1) - f_n(M_2)]^2}{\sigma_{n,1}^2 + \sigma_{n,2}^2}$$

where $\sigma_{n,j}$ is the inferred measurement error associated with data point n in model j, and N_{XL} is the total number of crosslinks. A cutoff of 0.05 was used. In multi-state modeling, data-based clustering is preferred to structure-based clustering (*e.g.,* using Cα-RMSD as the distance metric) because it reflects the degeneracy of models that would generate the same data and because it provides a natural way of mixing X_i, w_i, and σ_n^E that is not possible in structure-based clustering. Because the sample is drawn from the posterior distribution, the cluster population is proportional to the average posterior probability of its members. We focused our analysis on the clusters with a population greater than 3%. The structural model precision of a given cluster was defined as the median of the RMSD distribution calculated on all pairs of cluster members.

3.5.8 – Quantitative Structural Analysis

We gathered structures from two-component systems with multiple structures of the same domain, listed in **Table 5**. For each domain, we define the bundle axis by first selecting two pairs of equivalent residues, one from each chain, calculating the α-carbon to α-carbon vector between those two residues for both chains, and then summing these two vectors to create the axis vector. We define the bundle axis vector for one structure (the first PDB ID in each table row) to be the z axis, arbitrarily specify an x axis orthogonal to the z axis, and define the y axis perpendicular to x and z, using a right-handed coordinate system. We then align the remaining domains to the first structure using CEAlign [86] along the domain boundaries listed in **Table 6**.

Table 6 - Parameters used for domain fitting.

Domain	Protein	Organism	Domain Boundary	Helix Residues (Aligning Residues)	Aligning Residues
Sensor domain	Tar	*S. typhimurium*	42-174	42-57, 155-174	50,61
Sensor domain	CitA	*K. pneumoniae*	6-129	12-25,45-51	15,23
HAMP	Af1503	*A. fulgidus*	279-328	280-297,310-328	284,295
DHp	EnvZ	*S. flexneri*	335-385	333-345,373-385	383-376
DHp	DesK	*B. subtilis*	191-232	182-198,224-238	187,198

We fit each helix to a straight ideal helix (2.3 Å α-carbon radius, 3.6 residues/turn, 1.5 Å rise/residue) and extract six geometric parameters that define the helix's position and orientation by fitting a sequence of six motions (**Figure 12**) using the Levenberg-Marquardt algorithm from the GNU scientific library [29] (10^{-4} absolute tolerance, 10^{-4} relative tolerance, maximum 20 iterations). For each helix and for all six motions, we measured the maximum variation (range) of the fitted parameter. We normalize all rotations to distances by converting degrees to subtended arcs using a radius equivalent to the distance of an ideal β-carbon at a helix endpoint from the focal point of the rotation. This corresponds to an arc radius of 4 Å for rotations about the helix axis and an arc radius of (1.5 Å x (# of helix residues) / 2) for tilting motions. The full set of calculated displacements is given in **Figure 15**.

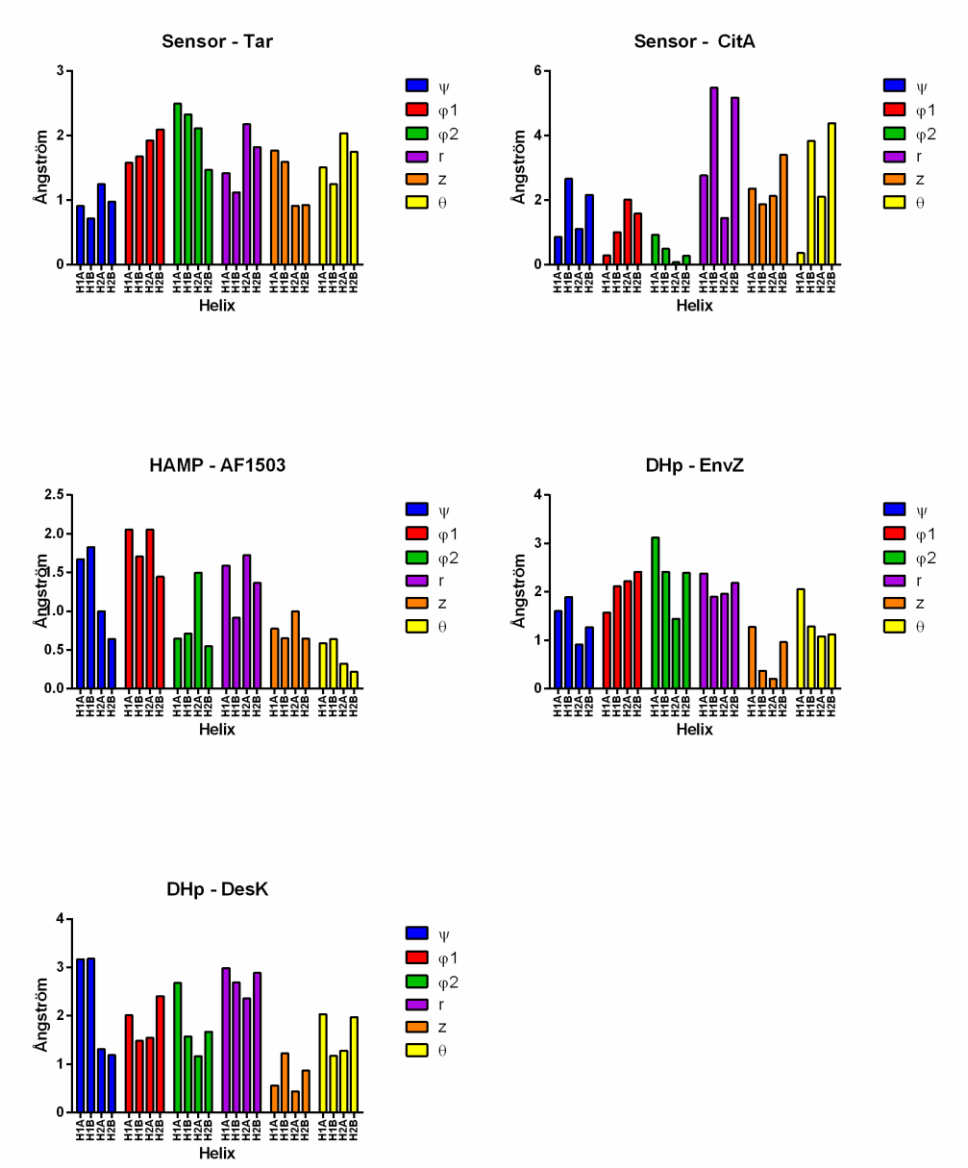

Figure 15 - Measured differences between equivalent helices in two component systems. Each chart corresponds to a single set of domains referenced by domain type (e.g. Sensor or HAMP) and protein (e.g. AF1503 or EnvZ). Each chart groups the differences by direction, with four measured differences per direction, one for each helix in a four-helix bundle: H1A) Helix 1 - Chain A, H1B) Helix 1 – Chain B, H2A) Helix 2 – Chain A, H2B) Helix 2 – Chain B. All measurements are calculated maximum displacements along each degree of freedom for an ideal β-carbon at a helix endpoint.

3.10 - Acknowledgments

Graham D Clinthorne, Massimiliano Bonomi, Riccardo Pellarin, Kathleen S Molnar, Gabriel Gonzalez**,** Shalom D Goldberg, Andrej Sali, and William F DeGrado are coauthors on this manuscript. Graham D Clinthorne and Kathleen S Molnar contributed disulfide crosslinking data, Massimiliano Bonomi and Riccardo Pellarin performed Bayesian modeling based on the data. Gabriel Gonzalez performed the quantitative structural analysis. All authors were involved in writing the manuscript.

CHAPTER 4 – Discussion

I approached the problem of connectable protein design using as few simplifying constraints as possible. This means that I explored solutions that did not impose symmetry on the protein, did not limit designs to pure assemblies of α-helices, and did not assume a minimum size scale of interest. All of these assumptions would have simplified the problem, but would have greatly narrowed the applicability of connectable protein design.

Connectable protein design moves beyond simply cataloguing designable motifs, but instead focus on tools and principles for joining them together to generate correct protein structures with as few conflicts as possible. This would then combine the best of two worlds: the reliability and robustness of battle-tested natural protein components with the novelty and flexibility of *de novo* protein design.

In Chapter 2, the Suns atomic search engine addressed connectability on the atomic scale, where chemical and steric constraints are precise and there is very little leeway to engineer in flexible and reusable interfaces. In Chapter 3, I studied connectability on the protein domain scale by structurally analyzing two-component systems to understand how they implement reusable and modular signal transduction interfaces.

4.1 – Connecting designable atomic substructures

The Suns search engine provides the first method for interactively connecting designable atomic fragments together into a larger whole. This greatly improves the utility of these atomic motifs, which are small, inflexible, and difficult to customize. Without Suns a protein designer would

have great difficulty discovering other compatible motifs that can be correctly integrated into an existing design while satisfying all covalent, electrostatic, and steric constraints.

Fragments built from search queries automatically get internal bond lengths and angles correct and avoid internal steric clashes since they all derive from natural protein structures. Even larger protein fragments pieced together from many small searches, as in **Figure 3**, one can still preserve this conflict-free property. However, this is not automatic; when searching for a designable motif to fill a spatial region, one must carefully select search queries that include existing chemical motifs likely to sterically or electrostatically interact with that region.

Search results also return a extra information that can guide the user and inform the connection process, since they preserve the immediate surroundings of all matches. This allows the user to discover fortuitous electrostatic interactions, supporting hydrogen bond networks, and hydrophobic packing interactions that were not part of their original search query. These are the kinds of useful interactions that an automated connection workflow might ignore, but that a human might find meaningful.

The Suns search engine makes this iterative connection workflow feasible through improved search speed and careful integration with molecular graphics software. Without these two improvements the time investment of connectability-based design becomes prohibitive at the atomic level.

4.1.1 – Mixed initiative

The Suns search engine is the first tool that mixes human intelligence with machine intelligence when designing interactions at the atomic scale. This allows a human to engage in a "dialog" with the computer throughout the entire process. This is known as a "mixed initiative" [43], where man and machine switch back and forth between taking the initiative, and contrasts with the traditional division of labor in protein design where the high-level specification is usually done entirely by humans and then handed off to a computer to implement. This division of labor leaves no opportunities for the computer to assist the human in the high-level design of the protein nor does it permit the human to assist the computer in the discovery of solutions.

The Suns search engine breaks down this barrier in both directions. The machine can now open up a window to the Protein Data Bank for the user, digesting large numbers of protein structures to provide a visual summary suitable for human consumption. This informs the user what designable interactions are available and compatible with their current blueprints, informing the specification process. In the other direction, Suns lets the human guide the implementation process by selecting what permutations of motifs to connect and test. This takes advantage of the human mind's power to deftly explore large, combinatorial solution spaces using efficient heuristics, in the same way that FoldIt uses human ingenuity to solve protein folding more efficiently [51].

Suns also blurs the division between specification and implementation. The traditional protein design process emphasizes a "waterfall" approach where information flows in one direction from specification to implementation. In contrast, Suns lets the protein designer interleave the

specification and implementation, discovering new motifs to incorporate as they virtually grow their protein.

4.1.2 – Importance of speed and interactivity

The Suns search engine introduces the notion that greater speed is itself a source of scientific novelty. When a tool takes hours to use then users will contemplate workflows that use such a tool once. If a tool takes seconds to use then users can begin to experiment with new workflows that invoke the tool repeatedly with constant iteration and feedback.

For example, if the atomic search pipeline could be improved to be even more efficient and compute all results within 16 milliseconds or less then this would open up the possibility of concurrently viewing search results in real-time while manipulating a protein structure. Clusters of potential matches would appear or disappear fluidly as the user continuously varied a bond length or angle. Similar improvements to secondary or tertiary structure search would allow one to view preferred backbone clusters fade in and out of view while experimenting with different orientations of one helix packing against another helix.

Innovations in the Suns search engine could potentially be used to improve the speed of the MaDCaT search engine, such as tokenizing searchable elements. Speed gains in MaDCaT would encourage an interactive workflow when studying secondary or tertiary structure, improving the ease of connecting designable interactions on a larger length scale. Additionally, integrating MaDCaT with PyMOL would allow users to seamlessly switch between both Suns and MaDCaT, promoting easier flow of information between both programs.

4.2 – Connecting protein domains

Protein domains provide more opportunities for engineering loosely coupled and flexible interfaces, because there is more material to work with and customize. Two-component systems exemplify this interfacial flexibility, where diverse domains can be assembled in varying permutations to generate functional signal transduction chains. Scientists have successfully recreated chimeras between two-component systems and chemoreceptors that signal correctly, such as the Tar-EnvZ [96], Trg-EnvZ [7], Tar-Tap [100], and Dcu-EnvZ [31] hybrids, which is a testament to how modular they are.

These loose couplings are made possible by four helix bundles that repack in a motion resembling scissoring, where one pair of opposing helices moves inward and the other pair moves outward. In PhoQ, we observe that the transition between the transmembrane domain and the HAMP domain is not a single continuous helix, yet still transmits the signal correctly. This suggests that perhaps scissoring preserves interfacial flexibility and loose coupling by not relying on the presence of a continuous helix bridging signal transduction domains.

However, I do not mean loose coupling in the sense of structural flexibility. Zhu and Inouye demonstrated that the interface between two-component domains can be very sensitive to changes in helical phase caused by inserting or deleting a single residue [107]. Instead, I use loose coupling in the sense of modularity, that we can splice in new components at well-defined junction points (such as regions of high sequence homology) to generate novel permutations of function.

4.2.1 – Signal transduction by helix bundle repacking

Two-component signal transduction highlights how repacking of four-helix bundles can be used as a mechanism for introducing structural variability within protein interfaces in a controlled manner. Scissoring motions within a four-helix bundle limit helices to move along large and predictable trajectories towards or away from the bundle axis, making them ideal candidates for transmitting conformational change along otherwise flexible connections.

These repacking motions can be quite large, on the scale of 3 Å. This might make them less sensitive to noise from background structural fluctuations, which usually exhibit root-mean-square fluctuations of 1 Å or less [74]. This may be a general principle for high-fidelity signal transduction across domain boundaries, where transmitted motions must be large to avoid spurious signaling events.

4.3 – Multiscale, connectable protein design

Throughout this thesis I explore a "multiscale" approach to connecting components, which mixes together solutions at different length scales. For atomic-scale protein design I built a protein search engine in order to discover and incorporate atomic substructures with correct chemistry, ideal electrostatic interactions, and no steric clash. For protein design on the secondary or tertiary structure scale I combine the Suns search engine with MaDCaT to discover host scaffolds compatible with smaller substructures built using Suns. On an even larger scale I approach the problem from a different angle and studied reusable natural domains from two-component systems which are sufficiently large to accommodate modular and flexible interfaces.

Designing at different length scales can be thought of as the protein design analog of multiscale modeling of biological systems [94]. We incorporate information at several different length scales simultaneously: the choice of secondary structure might impact our choice of possible atomic substructures, and, vice versa, our choice of atomic structure might feed back into our choice of secondary structure. Similarly, both would affect scaffold scaffold selection.

For multiscale protein design to work there must be a clear way for information to flow between different length scales, just as in multiscale modeling. In Chapter 2, my thesis explored one such flow of information by taking small fragments generated by Suns and coarse-graining them into α-carbon traces suitable for MaDCaT searches. The next step would be to explore if we could efficiently transition in the opposite direction by first selecting a scaffold using a MaDCaT search and then switching to Suns by converting α-carbon traces to atomic representations using Scwrl [56] and then using Suns to refine and locally mutate key substructures of interest.

There must also be a bridge between atomic / secondary structure to tertiary structure. This can be solved by the curation of protein domains that express well, fold robustly, and crystallize easily. Then Suns or MaDCaT searches could restrict themselves to matches from this curated set to greatly ease the experimental component of protein design. The David Baker research group has already made progress in this area by curating a set of domains that successfully express in *E. coli*, form monomers, and lack disulfide bonds [28], and these could form the basis of such a curated set.

All of these minor transitions and barriers switching between diverse tools should be removed before we can truly consider the multi-scale design problem solved. This calls for better

integration of scientific software with molecular graphics software to facilitate cross-pollination of design methodologies.

These approaches still need to be validated experimentally to see if reusable components identified this way fold as predicted, even when connected with unnatural partners. This technique has been proven for designable tertiary interactions found using MaDCaT [105] , but has not yet been proven for designable atomic level interactions generated by Suns. Experimentally establishing that designability works on the atomic scale would open up low-level protein design to a much broader and audience of non-scientific and non-technical users.

BIBLIOGRAPHY

1. Abràmoff MD, Magalhães PJ, Ram SJ (2004) Image processing with ImageJ. Biophotonics international 11: 36-42.

2. Albanesi D, Martín M, Trajtenberg F, Mansilla MC, Haouz A, et al. (2009) Structural plasticity and catalysis regulation of a thermosensor histidine kinase. Proceedings of the National Academy of Sciences of the United States of America 106: 16185-16190.

3. Alber F, Dokudovskaya S, Veenhoff L, Zhang W, Kipper J, et al. (2007) The molecular architecture of the nuclear pore complex. Nature 450: 695-701.

4. Alberts B, Johnson A, Lewis J, Raff M, Roberts K (2002) Molecular biology of the cell 4th edition.

5. Arnold FH (1998) Design by directed evolution. Accounts of chemical research 31: 125-131.

6. Barnakov A, Altenbach C, Barnakova L, Hubbell WL, Hazelbauer GL (2002) Site-directed spin labeling of a bacterial chemoreceptor reveals a dynamic, loosely packed transmembrane domain. Protein science: a publication of the Protein Society 11: 1472-1481.

7. Baumgartner JW, Kim C, Brissette R, Inouye M, Park C, et al. (1994) Transmembrane signalling by a hybrid protein: communication from the domain of chemoreceptor Trg that recognizes sugar-binding proteins to the kinase/phosphatase domain of osmosensor EnvZ. Journal of bacteriology 176: 1157-1163.

8. Bommarius AS, Blum JK, Abrahamson MJ (2011) Status of protein engineering for biocatalysts: how to design an industrially useful biocatalyst. Current opinion in chemical biology 15: 194-200.

9. Bonomi M, Pellarin R, Spill Y, Nilges M, DeGrado WF, et al. (2013) Modeling multiple structural states of macromolecules based on cysteine cross-linking. In preparation.

10. Brennan DJ, O'Connor DP, Rexhepaj E, Ponten F, Gallagher WM (2010) Antibody-based proteomics: fast-tracking molecular diagnostics in oncology. Nature Reviews Cancer 10: 605-617.

11. Brin S, Page L (1998) The anatomy of a large-scale hypertextual Web search engine. Computer networks and ISDN systems 30: 107-117.

12. Canutescu AA, Shelenkov AA, Dunbrack RL (2003) A graph-theory algorithm for rapid protein side-chain prediction. Protein science 12: 2001-2014.

13. Carter PJ (2011) Introduction to current and future protein therapeutics: a protein engineering perspective. Experimental cell research 317: 1261-1269.

14. Casino P, Rubio V, Marina A (2009) Structural insight into partner specificity and phosphoryl transfer in two-component signal transduction. Cell 139: 325-336.

15. Chen WW, Shakhnovich EI (2005) Lessons from the design of a novel atomic potential for protein folding. Protein science 14: 1741-1752.

16. Chervitz SA, Falke JJ (1996) Molecular mechanism of transmembrane signaling by the aspartate receptor: a model. Proceedings of the National Academy of Sciences of the United States of America 93: 2545-2550.

17. Cheung J, Bingman CA, Reyngold M, Hendrickson WA, Waldburger CD (2008) Crystal structure of a functional dimer of the PhoQ sensor domain. Journal of Biological Chemistry 283.

18. Claessen K (2004) Parallel parsing processes. Journal of Functional Programming 14: 741-757.

19. Dago AE, Schug A, Procaccini A, Hoch JA, Weigt M, et al. (2012) Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. Proceedings of the National Academy of Sciences 109: E1733-E1742.

20. Dahiyat BI, Mayo SL (1997) De novo protein design: fully automated sequence selection. Science 278: 82-87.

21. Daura X, Gademann K, Jaun B, Seebach D, van Gunsteren WF, et al. (1999) Peptide folding: when simulation meets experiment. Angewandte Chemie International Edition 38: 236-240.

22. Diensthuber RP, Bommer M, Gleichmann T, Möglich A (2013) Full-Length Structure of a Sensor Histidine Kinase Pinpoints Coaxial Coiled Coils as Signal Transducers and Modulators. Structure.

23. Doolittle JM, Gomez SM (2011) Mapping protein interactions between Dengue virus and its human and insect hosts. PLoS Negl Trop Dis 5: e954.

24. Dunin-Horkawicz S, Lupas AN (2010) Comprehensive analysis of HAMP domains: implications for transmembrane signal transduction. Journal of molecular biology 397: 1156-1174.

25. Dutta R, Qin L, Inouye M (1999) Histidine kinases: diversity of domain organization. Mol Microbiol 34: 633-640.

26. Falke JJ, Erbse AH (2009) The piston rises again. Structure (London, England: 1993) 17: 1149-1151.

27. Ferris Hedda U, Dunin-Horkawicz S, Hornig N, Hulko M, Martin J, et al. (2012) Mechanism of Regulation of Receptor Histidine Kinases. Structure 20: 56-66.

28. Fleishman SJ, Whitehead TA, Ekiert DC, Dreyfus C, Corn JE, et al. (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science 332: 816-821.

29. Galassi M, Gough B (2005) GNU scientific library: reference manual: Network Theory.

30. Galperin MY, Nikolskaya AN, Koonin EV (2001) Novel domains of the prokaryotic two-component signal transduction systems. FEMS Microbiology Letters 203: 11-21.

31. Ganesh I, Ravikumar S, Lee SH, Park SJ, Hong SH (2013) Engineered fumarate sensing Escherichia coli based on novel chimeric two-component system. Journal of biotechnology.

32. Gao R, Lynn DG (2005) Environmental pH sensing: resolving the VirA/VirG two-component system inputs for Agrobacterium pathogenesis. Journal of bacteriology 187: 2182-2189.

33. García Véscovi E, Soncini FC, Groisman EA (1996) Mg2+ as an extracellular signal: environmental regulation of Salmonella virulence. Cell 84: 165-174.

34. Goldberg SD, Clinthorne GD, Goulian M, DeGrado WF (2010) Transmembrane polar interactions are required for signaling in the Escherichia coli sensor kinase PhoQ. Proceedings of the National Academy of Sciences of the United States of America 107: 8141-8146.

35. Goldberg SD, Soto CS, Waldburger CD, DeGrado WF (2008) Determination of the Physiological Dimer Interface of the PhoQ Sensor Domain. Journal of Molecular Biology.

36. Goldsmith M, Tawfik DS (2012) Directed enzyme evolution: beyond the low-hanging fruit. Current Opinion in Structural Biology 22: 406-412.

37. Gordeliy VI, Labahn J, Moukhametzianov R, Efremov R, Granzin J, et al. (2002) Molecular basis of transmembrane signalling by sensory rhodopsin II–transducer complex. Nature 419: 484-487.

38. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, et al. (2003) Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. Journal of molecular biology 331: 281-299.

39. Grigoryan G, Kim YH, Acharya R, Axelrod K, Jain RM, et al. (2011) Computational design of virus-like protein assemblies on carbon nanotube surfaces. Science 332: 1071-1076.

40. Habeck M, Rieping W, Nilges M (2006) Weighting of experimental evidence in macromolecular structure determination. Proceedings of the National Academy of Sciences of the United States of America 103: 1756-1761.

41. Holm L, Sander C (1993) Protein structure comparison by alignment of distance matrices. J Mol Biol 233: 123-138.

42. Holm L, Sander C (1994) The FSSP database of structurally aligned protein fold families. Nucleic Acids Res 22: 3600-3609.

43. Horvitz E. Principles of mixed-initiative user interfaces; 1999. ACM. pp. 159-166.

44. Hughes J (1989) Why functional programming matters. The computer journal 32: 98-107.

45. Hulko M, Berndt F, Gruber M, Linder JU, Truffault V, et al. (2006) The HAMP Domain Structure Implies Helix Rotation in Transmembrane Signaling. Cell 126: 929-940.

46. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. Journal of molecular graphics 14: 33-38.

47. Hutton G, Meijer E (1998) Monadic parsing in Haskell. Journal of functional programming 8: 437-444.

48. Kabsch W (1976) A solution for the best rotation to relate two sets of vectors. Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography 32: 922-923.

49. Kaspar S, Perozzo R, Reinelt S, Meyer M, Pfister K, et al. (1999) The periplasmic domain of the histidine autokinase CitA functions as a highly specific citrate receptor. Molecular microbiology 33: 858-872.

50. Keasling JD (2010) Manufacturing molecules through metabolic engineering. Science 330: 1355-1358.

51. Khatib F, Cooper S, Tyka MD, Xu K, Makedon I, et al. (2011) Algorithm discovery by protein folding game players. Proceedings of the National Academy of Sciences 108: 18949-18953.

52. Kindrachuk J, Paur N, Reiman C, Scruten E, Napper S (2007) The PhoQ-Activating Potential of Antimicrobial Peptides Contributes to Antimicrobial Efficacy and Is Predictive of the Induction of Bacterial Resistance? Antimicrobial Agents and Chemotherapy 51: 4374-4381.

53. King NP, Sheffler W, Sawaya MR, Vollmar BS, Sumida JP, et al. (2012) Computational design of self-assembling protein nanomaterials with atomic level accuracy. Science 336: 1171-1174.

54. Korendovych IV, Kulp DW, Wu Y, Cheng H, Roder H, et al. (2011) Design of a switchable eliminase. Proceedings of the National Academy of Sciences 108: 6823-6827.

55. Kortemme T, Baker D (2004) Computational design of protein–protein interactions. Current opinion in chemical biology 8: 91-97.

56. Krivov GG, Shapovalov MV, Dunbrack RL (2009) Improved prediction of protein side-chain conformations with SCWRL4. Proteins: Structure, Function, and Bioinformatics 77: 778-795.

57. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, et al. (2003) Design of a novel globular protein fold with atomic-level accuracy. Science 302: 1364-1368.

58. Lanci CJ, MacDermaid CM, Kang S-g, Acharya R, North B, et al. (2012) Computational design of a protein crystal. Proceedings of the National Academy of Sciences 109: 7304-7309.

59. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. Journal of applied crystallography 26: 283-291.

60. Lee AI, Delgado A, Gunsalus RP (1999) Signal-Dependent Phosphorylation of the Membrane-Bound NarX Two-Component Sensor-Transmitter Protein ofEscherichia coli: Nitrate Elicits a Superior Anion Ligand Response Compared to Nitrite. Journal of bacteriology 181: 5309-5316.

61. Lemmin T, Soto CS, Clinthorne G, DeGrado WF, Dal Peraro M (2013) Assembly of the Transmembrane Domain of E. coli PhoQ Histidine Kinase: Implications for Signal Transduction from Molecular Simulations. PLoS computational biology 9: e1002878.

62. Li H, Helling R, Tang C, Wingreen N (1996) Emergence of preferred structures in a simple model of protein folding. SCIENCE: 666-669.

63. Lynch BA, Koshland D (1991) Disulfide cross-linking studies of the transmembrane regions of the aspartate sensory receptor of Escherichia coli. Proceedings of the National Academy of Sciences 88: 10402-10406.

64. Margeot A, Hahn-Hagerdal B, Edlund M, Slade R, Monot F (2009) New improvements for lignocellulosic ethanol. Current opinion in biotechnology 20: 372-380.

65. Mascher T, Helmann JD, Unden G (2006) Stimulus Perception in Bacterial Signal-Transducing Histidine Kinases. Microbiology and Molecular Biology Reviews 70: 910-938.

66. Maslennikov I, Klammt C, Hwang E, Kefala G, Okamura M, et al. (2010) Membrane domain structures of three classes of histidine kinase receptors by cell-free expression and rapid NMR analysis. Proceedings of the National Academy of Sciences of the United States of America 107: 10902-10907.

67. Metcalf DG, Kulp DW, Bennett JS, DeGrado WF (2009) Multiple approaches converge on the structure of the integrin αIIb/β3 transmembrane heterodimer. Journal of molecular biology 392: 1087-1101.

68. Miller JH (1972) Experiments in molecular genetics: Cold Spring Harbor Laboratory. 494 p.

69. Miller SI, Kukral AM, Mekalanos JJ (1989) A two-component regulatory system (phoP phoQ) controls Salmonella typhimurium virulence. Proceedings of the National Academy of Sciences of the United States of America 86: 5054-5058.

70. Moore JO, Hendrickson WA (2009) Structural analysis of sensor domains from the TMAO-responsive histidine kinase receptor TorS. Structure (London, England: 1993) 17: 1195-1204.

71. Morton GM (1966) A computer oriented geodetic data base and a new technique in file sequencing: International Business Machines Company.

72. Neidhardt FC, Bloch PL, Smith DF (1974) Culture medium for enterobacteria. Journal of bacteriology 119: 736-747.

73. Ness JE, Welch M, Giver L, Bueno M, Cherry JR, et al. (1999) DNA shuffling of subgenomic sequences of subtilisin. Nature biotechnology 17: 893-896.

74. Petsko GA, Ringe D (1984) Fluctuations in protein structure from X-ray diffraction. Annual review of biophysics and bioengineering 13: 331-371.

75. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, et al. (2004) UCSF Chimera--a visualization system for exploratory research and analysis. Journal of computational chemistry 25: 1605-1612.

76. Pontius J, Richelle J, Wodak SJ (1996) Deviations from standard atomic volumes as a quality measure for protein crystal structures. Journal of Molecular Biology 264: 121-136.

77. Prasad BV, Hardy ME, Dokland T, Bella J, Rossmann MG, et al. (1999) X-ray crystallographic structure of the Norwalk virus capsid. Science 286: 287-290.

78. Richardson JS, Richardson DC (1989) The de novo design of protein structures. Trends in biochemical sciences 14: 304-309.

79. Rieping W, Habeck M, Nilges M (2005) Inferential structure determination. Science 309: 303-306.

80. Robinson VL, Buckler DR, Stock AM (2000) A tale of two components: a novel kinase and a regulatory switch. Nature structural biology 7: 626-633.

81. Roy S, Aravind P, Madhurantakam C, Ghosh AK, Sankaranarayanan R, et al. (2009) Crystal structure of a fungal protease inhibitor from Antheraea mylitta. J Struct Biol 166: 79-87.

82. Royant A, Nollert P, Edman K, Neutze R, Landau EM, et al. (2001) X-ray structure of sensory rhodopsin II at 2.1-Å resolution. Proceedings of the National Academy of Sciences of the United States of America 98: 10131-10136.

83. Russel D, Lasker K, Webb B, Velazquez-Muriel J, Tjioe E, et al. (2012) Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. PLoS Biology 10: e1001244.

84. Schroeder U, Graff A, Buchmeier S, Rigler P, Silvan U, et al. (2009) Peptide nanoparticles serve as a powerful platform for the immunogenic display of poorly antigenic actin determinants. Journal of molecular biology 386: 1368-1381.

85. Senes A, Chadi DC, Law PB, Walters RFS, Nanda V, et al. (2007) E(z), a depth-dependent potential for assessing the energies of insertion of amino acid side-chains into membranes: derivation and applications to determining the orientation of transmembrane and interfacial helices. Journal of molecular biology 366: 436-448.

86. Shindyalov IN, Bourne PE (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein engineering 11: 739-747.

87. Shirvanyants D, Alexandrova AN, Dokholyan NV (2011) Rigid substructure search. Bioinformatics 27: 1327-1329.

88. Shyu C-R, Chi P-H, Scott G, Xu D (2004) ProteinDBS: a real-time retrieval system for protein structure comparison. Nucleic Acids Research 32: W572-W575.

89. Siegel JB, Zanghellini A, Lovick HM, Kiss G, Lambert AR, et al. (2010) Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction. Science 329: 309-313.

90. Sivia D, Skilling J (2006) Data analysis: a Bayesian tutorial.

91. Smith GP, Petrenko VA (1997) Phage display. Chemical reviews 97: 391-410.

92. Stock AM, Robinson VL, Goudreau PN (2000) Two-component signal transduction. Annual review of biochemistry 69: 183-215.

93. Taylor R, Kennard O, Versichel W (1984) The geometry of the N-HO= C hydrogen bond. 3. Hydrogen-bond distances and angles. Acta Crystallographica Section B: Structural Science 40: 280-288.

94. Telesco SE, Radhakrishnan R (2012) Structural systems biology and multiscale signaling models. Annals of biomedical engineering 40: 2295-2306.

95. Tzanov T, Calafell M, Guebitz GM, Cavaco-Paulo A (2001) Bio-preparation of cotton fabrics. Enzyme and Microbial Technology 29: 357-362.

96. Utsumi R, Brissette RE, Rampersaud A, Forst SA, Oosawa K, et al. (1989) Activation of bacterial porin gene expression by a chimeric signal transducer in response to aspartate. Science 245: 1246-1249.

97. Walters R, DeGrado W (2006) Helix-packing motifs in membrane proteins. Proceedings of the National Academy of Sciences 103: 13658-13663.

98. Wang C, Sang J, Wang J, Su M, Downey JS, et al. (2013) Mechanistic Insights Revealed by the Crystal Structure of a Histidine Kinase with Signal Transducer and Sensor Domains. PLoS biology 11: e1001493.

99. Wang G, Dunbrack RL, Jr. (2003) PISCES: a protein sequence culling server. Bioinformatics 19: 1589-1591.

100. Weerasuriya S, Schneider BM, Manson MD (1998) Chimeric chemoreceptors in Escherichia coli: signaling properties of Tar-Tap and Tap-Tar hybrids. Journal of bacteriology 180: 914-920.

101. Whitehead TA, Chevalier A, Song Y, Dreyfus C, Fleishman SJ, et al. (2012) Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. Nature biotechnology 30: 543-548.

102. Wu W, Hsiao SC, Carrico ZM, Francis MB (2009) Genome-Free Viral Capsids as Multivalent Carriers for Taxol Delivery. Angewandte Chemie International Edition 48: 9493-9497.

103. Xie W, Dickson C, Kwiatkowski W, Choe S (2010) Structure of the Cytoplasmic Segment of Histidine Kinase Receptor QseC, a Key Player in Bacterial Virulence. Protein and peptide letters 17: 1383.

104. Yohannan S, Faham S, Yang D, Whitelegge JP, Bowie JU (2004) The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors. Proceedings of the National Academy of Sciences of the United States of America 101: 959-963.

105. Zhang J, Grigoryan G (2013) Mining tertiary structural motifs for assessment of designability. Methods Enzymol 523: 21-40.

106. Zhao H, Arnold FH (1997) Functional and nonfunctional mutations distinguished by random recombination of homologous genes. Proceedings of the National Academy of Sciences 94: 7997-8000.

107. Zhu Y, Inouye M (2003) Analysis of the Role of the EnvZ Linker Region in Signal Transduction Using a Chimeric Tar/EnvZ Receptor Protein, Tez1. Journal of Biological Chemistry 278: 22812-22819.